Insights into enzyme allosteric inhibition mechanisms using computational studies

---

A thesis submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy in Chemistry

at the University of Canterbury

by

Eric Jean Maurice Lang

April 2016

Abstract

Allostery, the process by which action of an effector at one site of the protein provokes a functional response at a distant site, is critical in the regulation of metabolic pathways. Yet, despite its importance, allostery remains enigmatic and little detailed information is known about how the function of the protein is affected, how the allosteric signal is communicated or how allosteric regulation evolves in a protein.

The overall objective of this thesis is to provide new insights into the understanding of allostery by studying computationally allosteric regulation in the metabolic enzyme 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase (DAH7PS) which catalyses the first committed step in the biosynthesis of aromatic amino acids. Not only is DAH7PS a promising antimicrobial drug target, owing to its presence in microorganisms but not in animals, but also, the unique variety of allosteric regulation mechanisms found in this protein family makes it the ideal playground to study allostery.

Firstly, in Chapter 2, using a variety of sequence and phylogenetic analysis tools, the evolution of allosteric regulation in the DAH7PS family was explored. New subgroups of the DAH7PS family were identified and it was found that DAH7PS is likely to have acquired allostery by the recruitment of an N- or C-terminal regulatory domain via gene fusion. Loop

extensions and N-terminal domain mutations then gave rise to more complex allosteric regulation mechanisms.

In Chapter 3, a new computational method, which is based on the monitoring of calculated pK_a variations of ionisable residues over the course of molecular dynamics simulations, is used to identify allosteric communication pathways in protein relying on dynamic allosteric regulation. This approach is used to decipher the subtle allosteric regulation mechanism of DAH7PS from pathogen *Neisseria meningitidis*. A number of key charge-charge and hydrogen bond interactions were identified as being responsible for the communication of the allosteric signal in this protein. These predictions were verified experimentally and showed that pH variations can mimic the allosteric control, by changing the protonation state of the residues involved in the identified key interactions.

The allosteric regulation of the DAH7PS from hyperthermophile *Thermotoga maritima*, which relies on major domain motions, is explored using molecular dynamics simulations in Chapter 4. It was found that the flexibility of the allosteric domains impedes the entrance to the active site, explaining the lower catalytic activity observed experimentally when compared to a mutant lacking the regulatory domains. Furthermore, the simulations revealed that although the inhibition of the enzyme is achieved by blocking the access to the active site, the residual activity of the enzyme at high inhibitor concentrations observed experimentally, can be explained by the opening of a back door providing direct access to the catalytic site.

In Chapter 5, using a combination of X-ray crystallography, homology modelling and small-angle X-ray scattering, the allosteric regulation mechanism of DAH7PS from thermophilic soil bacterium *Geobacillus sp.* was studied. These methods revealed that binding

of the allosteric ligand to the allosteric regulatory domains stabilises a more compact conformation of the protein, limiting the catalytic functionality of the DAH7PS domain active site. In addition, molecular dynamics simulations are currently in progress to further explore the allosteric regulation of this enzyme are presented.

In Chapter 6, the different attempts to model the first step of the reaction catalysed by DAH7PS, using hybrid quantum mechanics / molecular mechanics simulations, are presented and recommendations for future work toward the complete modelling of the reaction are provided.

Finally, the implications of this work are discussed in the general context of allosteric regulation mechanisms.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 3

Lang EJM, Heyes LC, Jameson GB, Parker EJ: **Calculated pKa Variations Expose Dynamic Allosteric Communication Networks**. *Journal of the American Chemical Society* 2016, **138**:2036-2045

Please detail the nature and extent (%) of contribution by the candidate:

65%

Eric Lang designed the key methodlology and carried out the research. He wrote the majority of the manuscript, which was edited and contributed to by co-authors.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Emily Parker* Signature: Date: *6 April 2016*

Acknowledgements

First and foremost, I would like to express my most sincere gratitude to Professor Emily Parker for giving me this fantastic opportunity but also for being the most amazing supervisor I could have hoped for, leaving me the freedom to explore the areas I was interested in, providing me support, guidance and recommendations, making sure I had access to enough computational resources, facilitating collaborations and helping me to grow as a researcher. Thank you so much for everything Emily.

I am very grateful to Professor Adrian Mulholland, for hosting me for a month so I could gain hands-on experience in QM/MM modelling and for enabling me to access Bristol University supercomputer. Considerable thanks must go to Dr Marc Van der Kamp, who provided me with invaluable training in QM/MM modelling and shared his extensive experience with me while helping me running my own QM/MM calculations.

Thank you to Associate Professor Anthony Poole for guiding me through the abundance of bioinformatics tools and the diversity of practices that would have, otherwise, been daunting, and for providing me with critical recommendations on the analysis of my bioinformatics results.

Thank you to Dr Wanting Jiao for helping me with my first molecular dynamics (MD) simulations and providing me with useful recommendations, but also for rapidly leaving me to work on my own so I could gain confidence in running MD simulations and develop my own critical appraisal of the different MD techniques.

One of the aspect I enjoyed the most while working in the Parker group was the opportunity I had to collaborate and exchange ideas with experimental chemists and biochemists. Thanks to Dr Logan Heyes, for our fruitful collaboration on *Nme*DAH7PS and our extensive conversations on the interpretation of our results. Thanks to Yifei Fan for involving me in the preparation of the FRET experiments for *Gsp*DAH7PS and *Tma*DAH7PS, but also for our enjoyable chats about everything. Thanks to Yu Bai, for our collaboration on the SAXS analysis of *Gsp*DAH7PS and for our interesting discussions on so many subjects. Thanks to Thomas Cotton for explaining me some of the techniques used in the lab and for our entertaining exchanges. Thanks to Gert-Jan Moggre, for our interesting talks on QM modelling. Thanks to Jordyn Moore for the work on *Tma*DAH7PS. Thanks to Dr Penel Cross for useful discussions on *Tma*DAH7PS and on the analysis of SAXS experiments for *Gsp*DAH7PS. Thanks also to all the other past and present members of the Parker group, in particular, Dr Nicola Blackmore and Dr Leyla Bustamante for your support, but also Dr Ali Reza Nazmi, Dr Dmitri Joseph, Dr Tammie Cookson, and Sarah Wilson-Coutts.

Thank you to Professor Geoffrey Jameson for your wisdom, keen attention to details and critical comments during the writing of the two articles we co-authored. I am grateful to Dr Benjamin Roberts for the very useful recommendations on the minimisation, heating and equilibration protocols and on periodic images. I thank Dr Deborah Crittenden for her remarks and input. Thank you to Dr Jane Allison for our useful discussions.

I would like to thank the University of Canterbury High-Performance Computing (UCHPC) and New Zealand eScience Infrastructure (NeSI) teams for their support and for ensuring that everything ran smoothly. I thank in particular Dr François Bissey, Gene Soudlenkov, Dan Sun, Peter Maxwell, Jordi Blasco, Dr Celine Cattoen-Gilbert, Charles Bacon and Tony Dale. Thank you to Dr Sung Bae for writing the script that enables to restart automatically the MD simulations on UCHPC systems, and for helping me with the python scripts for the pK_a analysis.

I thank UCHPC for providing me with a doctoral scholarship to undertake my research as well as for the unlimited amount of core hours on UCHPC systems and for the travel fund. I also thank the Biomolecular Interaction Centre for providing me with a research scholarship for the last six months of my thesis and NeSI for allocating me more than 100,000 CPU/hours on the University of Auckland supercomputer.

Thank you to my family and my family-in-law for their support and encouragement from overseas via email and Skype.

Finally, to Sybile, thank you so much for having accepted to move to the other side of the world so I could live my dreams, for your constant support, for your encouragements, for believing in me, for your love, for our daughter. Merci mon Amour.

Lastly, I would like to dedicate this thesis to my daughter Eve. I hope this modest contribution to the general understanding of allostery and DAH7PS may be helpful to the development of new medicines so you can live in a safer, healthier world.

Table of contents

Abbreviations

ACT	Aspartate kinase, Chorismate mutase and TyrA
AM1	Austin Model 1
aMD	Accelerated molecular dynamics
Ape	*Aeropyrum pernix*
CM	Chorismate mutase
DAH7PS	3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase
DFT	Density functional theory
DFTB	Density functional based tight binding
E4P	Erythrose 4-phosphate
Eco	*Escherichia coli*
Ftu	*Francisella tularensis*
Gsp	*Geobacillus sp.*
HF	Hartree-Fock
HMM	Hidden Markov model
Hpy	*Helicobacter pylori*
ITC	Isothermal titration calorimetry
k_{cat}	Turnover number
KDO8PS	3-deoxy- D -*manno*-octulosonate 8-phosphate synthase
K_m	Michaelis-Menten constant
KNF	Koshland-Nemethy-Filmer

Lmo	*Listeria monocytogenes*
MD	Molecular dynamics
MM	Molecular mechanics
MP2	2nd order Møller-Plesset
MSA	Multiple sequence alignment
MSM	Markov state model
Mtu	*Mycobacterium tuberculosis*
MWC	Monod, Wyman and Changeux
Nme	*Neisseria meningitidis*
Pae	*Pseudomonas aeruginosa*
PDB ID	Protein Data Bank identification code
PEP	Phosphoenolpyruvate
Pfu	*Pyrococcus furiosus*
Pni	*Prevotella nigrescens*
QM	Quantum mechanics
QM/MM	Quantum mechanics/molecular mechanics
SAXS	Small-angle X-ray scattering
SCA	Statistical coupling analysis
Sce	*Saccharomyces cerevisiae*
SQM	Semi-empirical quantum mechanics
Tma	*Thermotoga maritima*
WHAM	Weighted histogram analysis method

List of figures

List of tables

Chapter 1 – Protein dynamics, allostery and molecular modelling

"Indeed the protein molecule model resulting from the X-ray crystallographic observations is a 'platonic' protein, well removed in its perfection from the kicking and screaming 'stochastic' molecule that we infer must exist in solution."

Gregorio Weber, Adv. Protein. Chem. 1975 [1]

1.1 Introduction

Proteins are dynamic entities. They are not static, rigid systems, an image that X-ray crystallography often conveys, and which is perpetuated in many biochemistry textbooks. Proteins move and, most importantly, their movements are related to function, such as in catalysis, signal transfer, or allosteric regulation. Allostery, the process by which an event at one site of the protein triggers a functional response at a distant site, is intimately linked to protein dynamics [2] and therefore to understand the allosteric regulation of a protein, one needs to study its dynamics. Several methods can be used to reveal functional motions involved in allosteric regulation, including but not limited to single-molecule Förster resonance energy

transfer (smFRET) [3], nuclear magnetic resonance (NMR) spectroscopy [4], hydrogen-deuterium exchange studied by mass spectroscopy [5], isothermal titration calorimetry (ITC) [6] and molecular dynamics (MD) simulations. In this chapter after introducing protein dynamics and the energy landscape theory, the concept of allostery will be explained both from a historical and a modern point of view. Then, the different methods enabling protein allostery to be studied computationally will be reviewed and illustrated with recent examples from the literature.

1.2 Protein dynamics

1.2.1 Historical perspective

Almost 60 years ago, the pioneering work of Linderstrøm-Lang on hydrogen-exchange in proteins led him to conclude that proteins were, in fact, dynamic systems and coined the term "*protein motility*" [7,8]. But it was only in the 1970's that initial experiments, including the use of fluorescence [9-11], hydrogen exchange studies [12], phosphorescence [13], NMR [14], and MD simulations [15], demonstrated the flexibility of proteins in solution [16-18]. From then, a whole body of experimental and computational studies have been accumulated, not only to verify that proteins are indeed dynamical entities, but also to demonstrate the inter-relationship between protein motions and function.

1.2.2 Energy landscapes and folding funnels

In a cornerstone paper published in 1991 [19], Frauenfelder, Silar and Wolynes introduced the concept of protein energy landscape, a rough high-dimensional surface that

describes the energy of the protein as a function of its conformational coordinates. Protein motions can be understood in terms of exploration of this energy landscape, where valleys correspond to the multitude of conformational states the protein can adopt and peaks to the energy barriers the protein has to cross in order to explore the next valley. The relative free energy difference between minima dictates the relative population of each conformational state based on a Boltzmann probability distribution, whereas the height of the energy barriers between states is linked to the rate of interconversion between conformations [20] and therefore the energy landscape theory describes protein dynamics both in thermodynamic and kinetic terms.

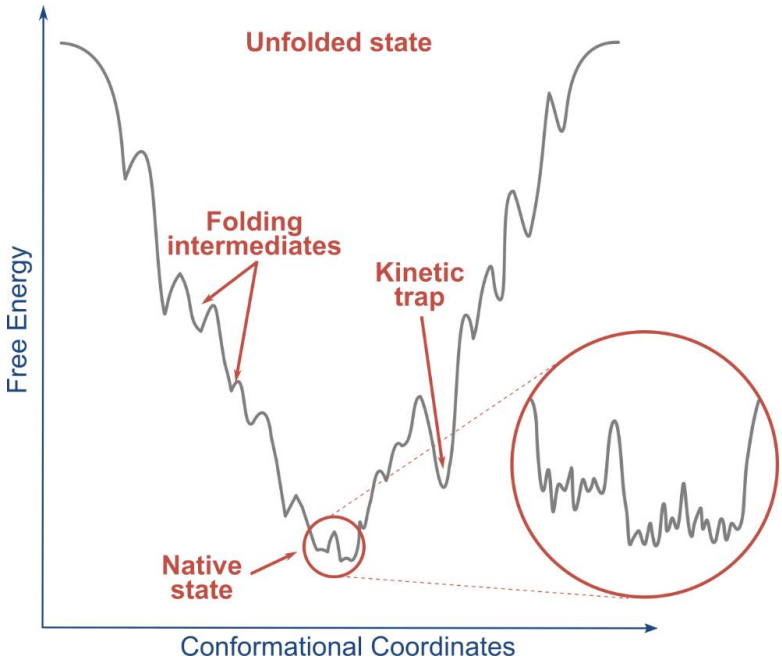


Figure 1.1. Schematic one-dimensional representation of a folding funnel. The unfolded protein on top is guided toward the bottom of the folding funnel (native state) thanks to the rugged funnel wall. Each local minimum along the path corresponds either to a folding intermediate or a kinetic trap. The native state, found at the bottom of the funnel, does not correspond to a single conformation but instead to the ensemble of conformations accessible to the folded protein.

The concept of the energy landscape was soon extended to include protein folding, leading to an energy landscape representation that resembles a rugged funnel (Figure 1.1), which describes the energy-biased folding pathway from a fully unfolded state at the top of the funnel to the native state at the bottom [21]. Protein folding can occur via multiple pathways [22], each being punctuated with local low energy states corresponding to folding intermediates or kinetic traps in the case these local minima are surrounded by high energy barriers the intermediate cannot easily escape. The bottom of the funnel does not correspond to a single well, but instead to a rugged portion of the energy landscape that describes the dynamics of the native state [23-25].

1.2.3 Taxonomy of the energy landscape

The portion of the energy landscape corresponding to the folded protein, which defines both amplitude and timescale of the motions accessible in the native state, can be divided into three tiers [19,20] as represented by the simplified one-dimensional energy landscape in Figure 1.2. Tier-1 and tier-2 dynamics correspond to localised fluctuations of small groups of atoms at the picosecond timescale (e.g. side-chain rotations) and nanosecond timescale (e.g. loop motions). Since the energy barriers between the states in tier-1 and 2 are small (inferior to 1 kT, k being the Boltzmann constant and T the temperature), conformational changes occur often and a large number of states are sampled. Tier-0 dynamics are characterised by the transition between a small number of states, which are separated by large energy barriers of several kT. Tier-0 dynamics correspond to broad domain motions, which require the collective movement of several groups of atoms and are therefore rare. Thus, although the transition itself is fast, the timescale of such transitions is important (microsecond to millisecond) due to their infrequency [26].

The energy landscape is not static but rather dynamic [25] as external factors such as ligand binding, changes in environmental conditions (pH, ionic concentration, etc.) or single point mutations will reshape the energy landscape [20,27,28], leading to differences in dynamics that can affect protein function both by changing the height of energy barriers and the relative energy difference between conformations. As we shall see in the following section, the effect of an allosteric effector can be fully explained by the associated reshaping of the energy landscape of the protein.

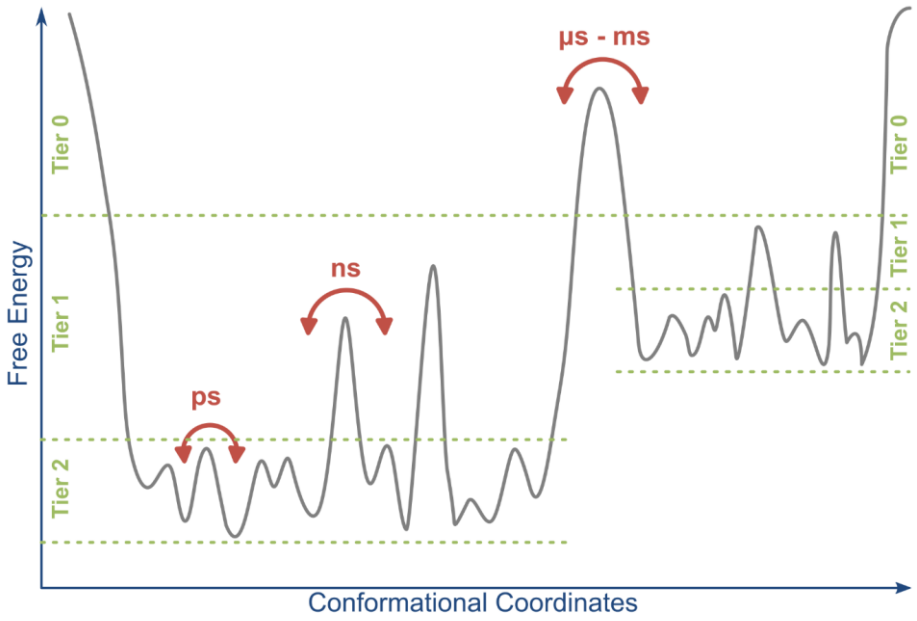


Figure 1.2. Simplified one-dimensional representation of a part of the energy landscape accessible to a folded protein. The free energy landscape can be decomposed into three tiers; tier-2 and tier-1 dynamics are localised fluctuations of small groups of atoms occurring at the picosecond and nanosecond timescale respectively, while tier-0 dynamics correspond to major collective motions of large groups of atoms and occur at the micro- to millisecond timescale. Adapted from [20].

1.3 The concept of allostery

1.3.1 The importance of allostery

Allostery is the process by which an effector alters the function of a biological macromolecule at one site by inducing a perturbation at a remote site. Allosteric regulation primarily affects proteins via the binding of a ligand (small molecules, lipids, RNA, DNA, proteins, etc.) at a site distant from a functional region, but can also arises from light excitation, changes in environmental conditions, mutations or covalent events such as phosphorylation or glycosylation [29].

Although allostery takes place at the protein level, its effect has implications at the cellular level [30], more specifically allostery serves as a mean to communicate environmental intra- and inter-cellular signals and is therefore able to alter specific cellular functions. For example, allostery is critical to the regulation of metabolic pathways by feedback regulating metabolic enzymes in response to the concentration of key intermediates or end-products [31]. Allostery is therefore of primary importance to cellular communication but because of this key role, misuse of allosteric regulation can lead to catastrophic effects on the cell. Aberrant allosteric regulation can lead to proteins being constantly switched on or off, not able to adequately respond to environmental factors. As cellular pathways are interconnected, uncontrolled protein functions can cascade into detrimental abnormal cell function typically leading to diseases [32].

Due to the key role of allostery in cellular processes [32], the benefits of developing drugs that target allosteric binding sites are twofold. Not only can they help to modulate rogue human protein function to restore a balanced cellular response, but they also may help to fight

pathogens by specifically disrupting bacterial or fungal metabolic pathways. However, understanding the mechanisms associated with allosteric regulation in proteins is a prerequisite for the development of effective allosteric medicines.

1.3.2 Historical perspective

In 1894, Fisher proposed the *lock-and-key* model [33] to explain the binding process of a ligand to a protein, based on the assumption that proteins were rigid bodies. The first attempt to introduce some flexibility in the protein was made 64 years later by Koshland, who proposed the *induced-fit* hypothesis [34], suggesting that a binding event induces conformational changes within the protein in order to accommodate for the ligand. This model was then extended to allostery: the Koshland-Nemethy-Filmer (KNF) induced fit model of allostery [35] proposes that conformational changes propagate through the protein matrix after binding of the ligand. In another approach, developed by Monod, Wyman and Changeux (MWC symmetrical model of allostery) [36,37], it is postulated that two conformations independent of ligand occupancy exist in equilibrium and that a ligand will stabilise the conformation to which it preferentially binds. The MWC model was however limited to symmetric multimeric proteins.

Both the MWC and KNF models present a number of limitations. Firstly, they are phenomenological in the sense that they do not answer at the atomic level of detail as to *how* the binding of a ligand at one site can affect the function of a distant site [38]. Moreover, both MWC and KNF original models were created based on the existence of only *two* distinct, well-defined conformations and the transition between them. In the case of the KNF model, each subunit can either adopt a conformation A or B (tertiary structural changes) whereas in the MWC model, a symmetrical multimeric unit can adopt either R or T state (corresponding to

quaternary structural changes). Consequently, both models are intrinsically linked to conformational changes between two clearly defined states. The only potential generalisation to allostery without conformational changes is mentioned in the MWC paper: *"It is also possible that the quaternary constraints might not force any significant sensu stricto "conformational" alteration of the protomers, but only, for example, a (symmetrical) redistribution of charge within the molecule."* [38].

It was only 20 years later, in 1984, that Cooper and Dryden [39] demonstrated that allostery could exist even in the absence of conformational changes via a purely dynamic, entropic process. Using a statistical thermodynamics framework, they showed that binding of an allosteric ligand can cause allosteric interaction energies of several kJ/mol via changes in frequency and amplitude of the protein atomic motions without changes in the mean atomic positions.

1.3.3 Today's view of allostery

The three major papers described above laid the foundations for the development and refinement of the current models of allostery. A large body of experimental, theoretical and computational work has permitted new insights into the modern understanding of allosteric regulation, which has been detailed in a number of excellent reviews [2,28,38,40,41]. Notwithstanding these advances, no real consensus has been reached and the mechanisms of allostery, especially at the atomic level, and these remain a hot topic still very much debated today.

In the following sections, the core concepts of allostery will be discussed in terms of energy landscape theory. Firstly, *conformational selection* and *induced fit*, which are based on the MWC and KNF models respectively (summarised in Figure 1.3A) will be described and then the concept of *dynamic allostery* will be presented. These three models correspond to the most extreme scenarios, the fact that real-life allosteric regulation mechanisms often consist as a combination of more than one of these models will then be explored.

1.3.3.1 Induced fit

The induced fit model of allostery corresponds to the KNF model and states that ligand binding induces a conformational change and that in the absence of ligand the protein remains in a single state. Figure 1.3B provides a simplified representation of the induced fit model in its original formulation. The protein exists in a single active conformation (global energy minimum), upon binding of the allosteric ligand, the energy landscape is entirely reshaped and the ligand-bound active conformation is now of high energy, the enzyme will therefore change its conformation to a more stable ligand-bound conformation, which in this case corresponds to the ligand-bound inactive enzyme.

This model might seem hardly compatible with what is now known on protein dynamics and the fact that the protein does not exist in a single conformation. However, one must remember the taxonomy of the energy landscape: if tier-0 is considered here, the induced fit is not incompatible with protein dynamics. Indeed, although the protein exists in a distinct tier-0 conformation, it does not prevent the exploration of the accessible subspace of the energy landscape, corresponding to the conformational substates from tier-1 and 2. The role of the

allosteric ligand, in this case, is to reshape the energy landscape in order to make favourable a tier-0 conformation so far inaccessible.

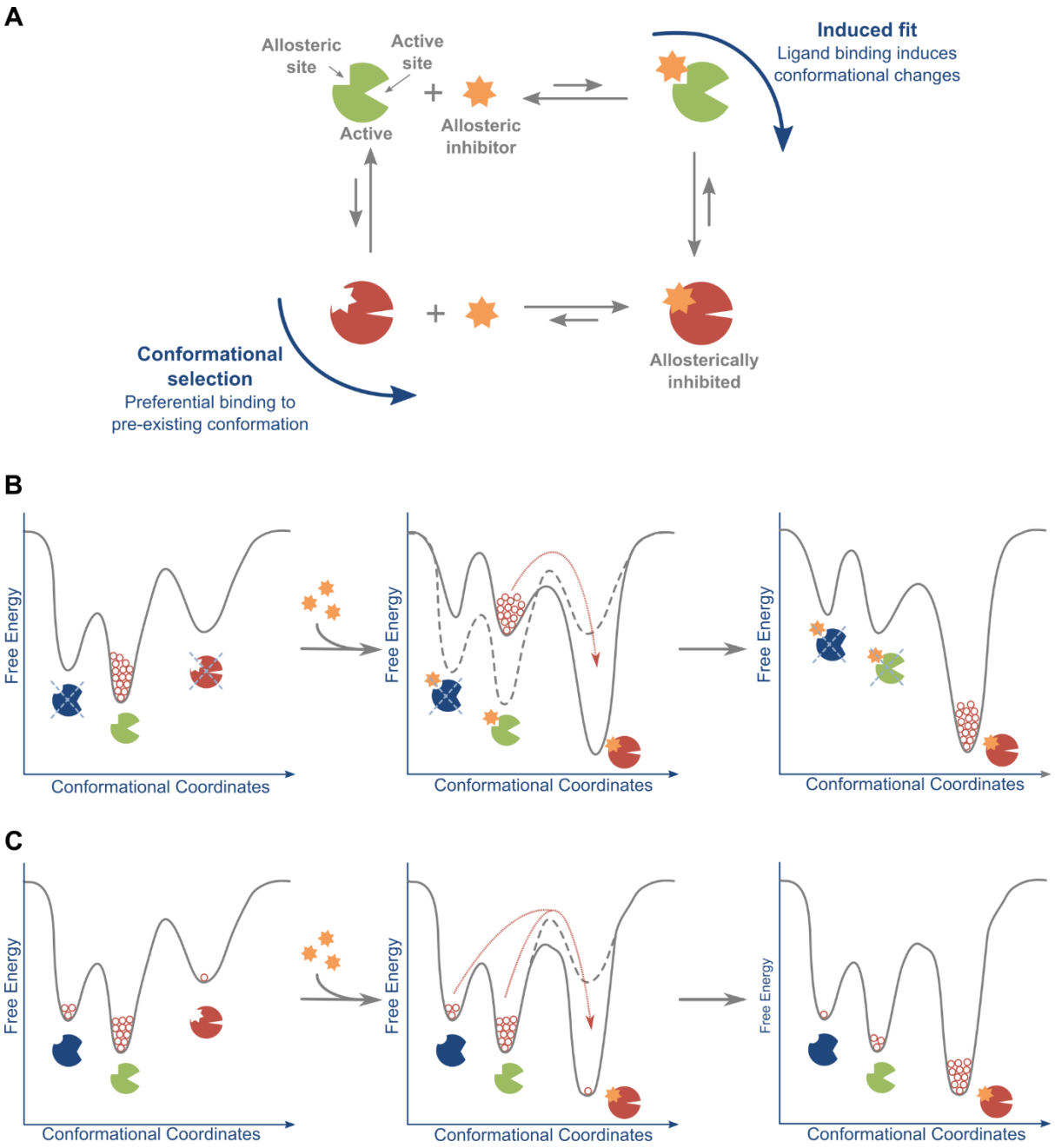


Figure 1.3. Induced fit versus conformational selection. A. Schematic representation of the induced fit versus conformational selection models. Whereas the induced fit model suggests that conformational changes in protein occur after the binding of a ligand, the conformational selection model considers the pre-existence of an ensemble of conformations sampled by the protein, the binding of a ligand occurs at a preferential conformation without

inducing any conformational changes. The active state of the protein is shown in green and the inactive form in red and the allosteric ligand as a yellow star. **B.** Schematic changes to a one-dimensional energy landscape associated with the induced fit model and **C.** the conformational selection model. The conformation of the enzyme corresponding to each energy minimum is represented on each diagram. Red corresponds to the inactive enzyme, green to the active enzyme and blue to another (less) active conformation. The beads in the energy basins represent the relative population of each conformation and the red arrows the population shifts. The energy surface prior to allosteric ligand binding is identified by a dashed line.

The induced fit model is therefore not incompatible with protein dynamics and in its present formulation should refer to a conformational change that is only possible in the presence of a ligand, without precluding any exploration of the energy landscapes accessible individually to the unbound and bound forms.

1.3.3.2 Conformational selection

The conformational selection model (or population shift model) is a reformulation of the MWC model developed essentially by Nussinov and Tsai that takes into account the free energy landscape theory [25,27,42]. Proteins in solution are in constant motion, sampling large ensembles of pre-existing conformations over time (and not only two as per the original MWC), the majority being weakly populated. The conformational selection model argues that an allosteric ligand will bind weakly populated, high energy conformations, altering the free-energy landscape in this region and shifting the population toward this bound conformation.

A simplified schematic of this model is presented in Figure 1.3C. Each protein conformation is sampled, but the relative population of each conformation is given by a Boltzmann distribution based on the conformations' relative free energy differences. The active

protein (green) being of lower energy, it corresponds to the most populated conformation, although higher energy states (blue and red) are also sampled. The addition of an allosteric ligand will lead to the preferential binding of the high energy inactive conformation (red). This region of the energy landscape will be affected and the ligand-bound inactive state becomes a global minimum. The population of the other conformations, now of higher energies, is therefore shifted toward the lowest energy state, leading to a population redistribution and the inactive conformation becoming the most populated one. The conformational selection model is applicable to virtually any tier of the energy landscape: for example, large domain motions (tier-0), loop motions (tier-1) or side chain reorientations (tier-2).

1.3.3.3 Dynamic allostery

Dynamic allostery as defined by Cooper and Dryden [39] is relatively recent when compared to the other two historical models, but perhaps more importantly, it seems to have been forgotten about for almost 20 years before molecular dynamics, NMR and ITC studies verified Cooper and Dryden's predictions (see for example [43,44]). The concept of dynamic allostery then gained traction and was discussed in great detail in two major reviews on the subject [2,45].

Dynamic allostery, which is entropically driven, arises when the mean position of atoms does not change in response to the allosteric signal, but the fluctuations around the mean do alter. These changes could be due to highly correlated low-frequency normal mode motions or random anharmonic fluctuations of individual atoms [39]. For example, let us consider an enzyme that requires large amplitude low-frequency hinge motions around its active site to effectively accept the substrates, perform the reaction and release the products. Binding of an

allosteric ligand might, for example, reduce the flexibility of this region of the enzyme, effectively leading to a stiffening of the active site (Figure 1.4A). In terms of the energy landscape, the result to be expected is a reduction of the "width" of the affected energy basin, accompanied with a lower energy to stabilise the new entity and raise the energy barriers to escape the energy well (Figure 1.4B) [46].

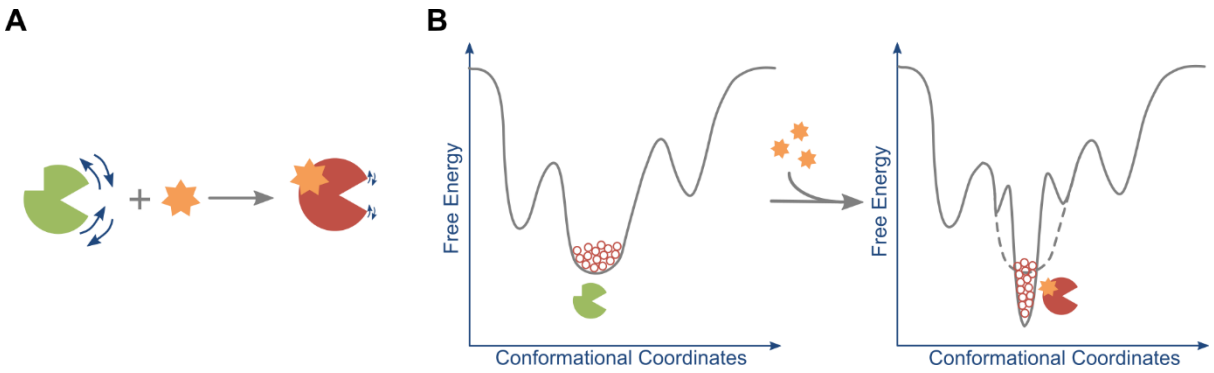


Figure 1.4. Dynamic allostery. A. The active enzyme (green) presents large fluctuations around the active site. These fluctuations are significantly reduced upon binding of an allosteric ligand (yellow star) rendering the enzyme inactive (red) without affecting the mean conformation of the enzyme. **B.** Schematic changes of a one-dimensional energy landscape corresponding to the dynamic allostery model. The beads in the energy basins represent the relative population of each conformation a. The surface prior to allosteric ligand binding is identified by a dashed line.

1.3.3.4 Combined mechanisms

The induced fit and conformational selection models have been the subject of decades of vivid disputes to determine once and for all which one is the "true" mechanism of allostery [47] (and binding as these two models primarily concern binding mechanisms), whereas dynamic allostery has only been considered more recently.

Although of the first two models, induced fit historically has been preferred [48], growing evidence in recent years supports the conformational selection mechanism as being the prevailing mechanism of allosteric regulation. Yet the subject remains highly controversial. In the following paragraph, a number of papers that highlight these discordances but that nevertheless tend to favour conformational selection over induced fit will be presented.

Vogt and Di Cera showed for example that conformational selection is sufficient to explain binding kinetics, whereas induced fit is only applicable in rare cases [49-51]. However, as adroitly noted by Weikl and Paul [52], again using kinetics, if the binding or unbinding events are faster than the time spent by the protein in different conformations, then the binding or unbinding events are decoupled from the conformational transition (they occur when the protein is in a given conformation) leading to a temporal ordering of the events (binding occurs either before or after the conformational transition). Based on the detailed balance principle, it signifies that the reverse pathway of the conformational transition actually is the induced fit mechanism and vice-versa (Figure 1.5A), making the two processes the two sides of the same coin. Vega and co-workers recently went one step further and tried to unify induced fit and conformational selection under a common model in which the equilibrium constant between conformations is the key parameter [53]. Although the idea is appealing, one might argue that in order to build this model the authors redefined induced fit as being an equilibrium between two states with one being weakly populated, whereas conformational selection was redefined as an equilibrium between two states, both being significantly populated. So effectively the authors replaced induced fit with conformational selection and the common model could essentially be interpreted as a unification between two types of conformational selection mechanisms. It has also been proposed that there is a continuum of binding mechanisms in which conformational selection and induced fit are the two extremes of this mechanism and,

depending on the transition rate between conformations and the ligand concentration, the fraction of induced fit versus conformational selection may vary [54,55]. One of the most well-accepted theories is that ligand binding may occur primarily via conformational selection, and then protein-ligand interactions could be optimised by an induced fit mechanism consisting of side-chains and backbone rearrangement, effectively leading to a mixed conformational selection – induced fit mechanism (Figure 1.5B). This compromise has been proposed as early as 2001 [56], and since then has received great interest [41,47,52,57].

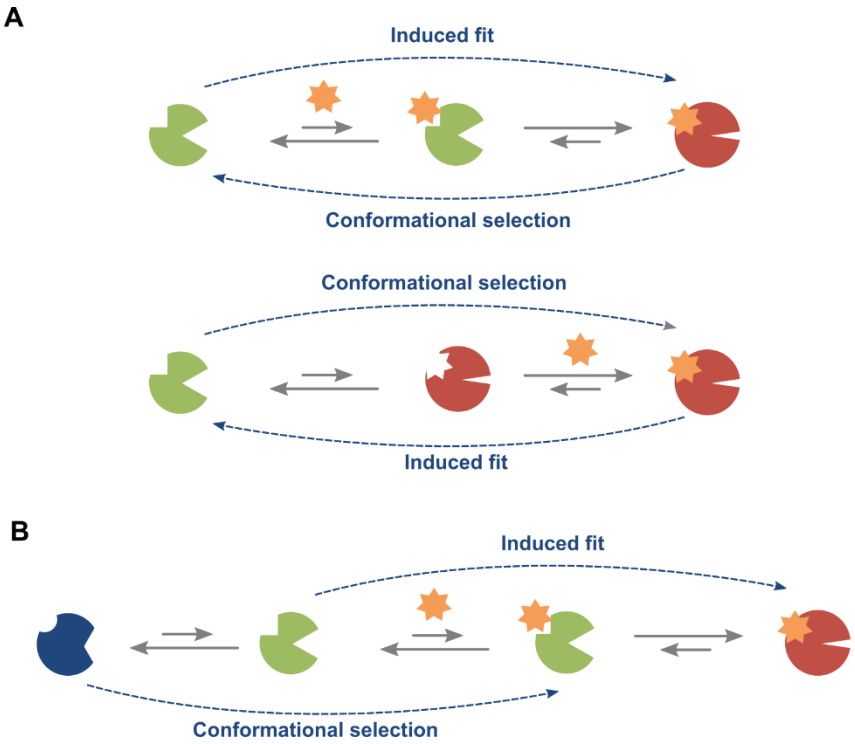


Figure 1.5. Combined mechanisms. A. Schematic description of Weilk and Paul [52] proposal applied here to allostery **B.** Schematic representation of a mixed conformational selection – induced fit allosteric mechanism where the allosteric ligand is considered not to bind the blue conformer, thus a conformational selection from blue to green is required permitting the ligand to bind, followed by an induced fit mechanism that increases the affinity of the binding site for the ligand.

Dynamic allostery can theoretically be seen as a mechanism taking place in the absence of conformational changes, in contrast with induced fit and conformational selection. However,

in practice this is not entirely the case, as the absence of conformational changes should not be interpreted literally and it would be better to refer to an absence of changes in shape or backbone conformation [45] or an absence of *major* conformational changes. Indeed, even Cooper and Dryden warned the reader that "*In practice, ligand-induced changes in* both *the mean conformation and dynamics are to be expected*" [39]. Moreover, failure to observe conformational changes does not necessarily imply that they are not at play, but it might instead signify a limitation of the experimental or computational technique used [58]. Therefore, protein dynamics do not preclude any structural changes and can even considered as part of mixed mechanisms involving changes in conformation and in dynamics as it has for example been nicely demonstrated in the case of Hsp70 [59].

1.3.3.5 Communication of the allosteric information

From the preceding discussions, a major question remains unanswered, how can the binding of a ligand at one site, affect a distant functional site? Or in other words, how does the allosteric signal communicated through a protein to perturb its functional site? To date, the answer to this question remains somewhat unresolved as well as protein-dependant. Although some authors have proposed a model in which allostery can exist in the absence of communication pathways [60,61], it is generally accepted that allosteric information is propagated through multiple, pre-existing pathways [29] across the protein matrix.

As early as 1985, it has been postulated that upon action of a specific effector, collective structural changes can propagate through the protein as a quake-like intramolecular motion named "proteinquake" [62]. Using femtosecond X-ray measurements, low-amplitude collective motions that spread throughout the protein at the picosecond timescale have been

observed [63,64], providing direct evidence of proteinquakes. These quake-like motions, that can involve partial unfolding of local regions of the protein, have been suggested as a possible mechanism of allosteric signal propagation [65].

Similarly, it has also been postulated that energy can flow through the protein matrix via specific channels connecting allosteric and functional sites [66] and the ballistic energy flow (i.e. a form of heat transport that occurs at the speed of sound without scattering and which can be extremely anisotropic) recently observed in the case of albumin [67] has been proposed as a mechanism for allosteric propagation of ligand binding energy through the protein structure.

Currently, one of the most popular ways to explain communication of the allosteric signal across a protein is to consider the protein as a network of residues interconnected via non-covalent interactions [68], such as hydrogen bonding, charge-charge interactions or hydrophobic interactions, but which could also involve conserved water molecules [69]. An allosteric effector can reorganise part of the network, thereby communicating allosteric information via local changes in conformation or dynamics arising from the formation and cleavage of the non-covalent interactions. This approach has led to the development of several methods to try to track down the atomistic mechanisms involved in communicating the allosteric signals [70].

1.3.3.6 Towards a general model of allostery?

Fifty years after the proposal of the KNF and MWC models, processes involved in allosteric regulation are still not completely understood and a large amount of work is still required before a consensus toward a grand unified theory of allostery is reached. Nonetheless,

two groups, in particular, have greatly contributed to improving our understanding of allostery via the proposal of general models of allostery: Nussinov and Tsai, who proposed the "*unified view*" of allostery [71,72], and Hilser, Wrabl and Motlagh who developed the *ensemble allosteric model* [28,61,73]. Both approaches seek to unify an extended MWC model and the concept of dynamic allostery with the energy landscape theory and both find their roots in a thermodynamical, macroscopic, description of allostery, although the unified view incorporates atomic-level structural elements in its approach.

1.3.4 Timescales relevant to allostery

In order to correctly characterise the dynamics of a protein, the three tiers need to be studied since biologically important conformational changes in a protein are often due to the subtle equilibrium between fast, localised fluctuations (tier-1 and 2 dynamics) and collective, slow motions (tier-0 dynamics).

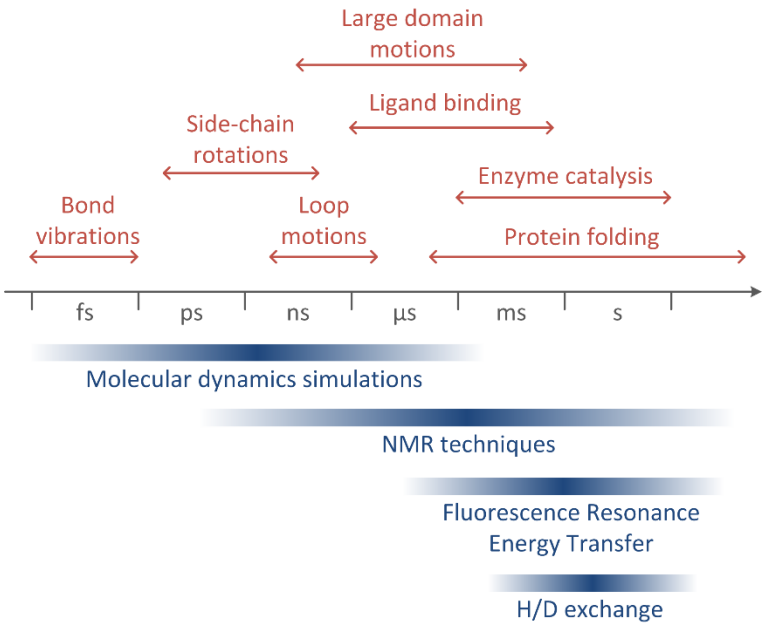


Figure 1.6. The timescale of protein dynamics and various biophysical techniques used to study the dynamics of allostery.

The timescale of protein dynamics spans several orders of magnitudes, from bond vibrations at the femtosecond-picosecond timescale to protein (un)folding events that could take place with time constants of hours [74]. Of particular relevance for allosteric regulation, local fluctuations such as side-chain rotations and loop motions correspond to a picosecond - nanosecond timescale, while highly correlated motions such as large domain motions and ligand binding occur at the microsecond to second timescale [20].

Among the number of techniques (Figure 1.6) that have been developed to follow protein dynamics involved in allosteric regulation, molecular dynamics (MD) simulations enable to access motions within the femtosecond to microsecond timescale and with atomic resolution. It is therefore a method of choice to study the allosteric regulation of proteins.

1.4 Studying protein allostery computationally

1.4.1 Exploring the conformational energy landscape

As we have seen, protein dynamics is intimately linked to allostery and to study allostery one must analyse protein movements. Molecular dynamics simulations are ideally suited for this task as they enable the study of protein dynamics at the picosecond – microsecond timescale at atomic resolution. Moreover, enhanced sampling methods make the observation of microseconds to milliseconds motions amenable within reasonable calculation times. Due to its atomic scale resolution, MD provides all the necessary information to identify allosteric

communication pathways and a number of methods have been developed to retrieve this information. Here, the theory behind MD will first be briefly described and a number of advanced sampling techniques relevant to allostery will be detailed and then methods enabling identification of allosteric communication pathways will be presented. Finally, alternative approaches to MD that permit an exploration of the energy landscape to be made will be briefly reviewed.

1.4.1.1 Classical all-atom molecular dynamics

MD simulation is one of the few techniques able to provide direct information on protein dynamics and has been referred to as the "computational microscope" of biochemistry [75]. Several books and reviews present MD in great detail and describe how this technique is implemented in a computationally efficient and robust manner (see for example [76-79]). Here, only the theoretical basis of MD and some of its variants will be summarised.

The central concept of MD simulation lies on Newton's second law:

$$\boldsymbol{F}_i = m_i \,.\, \boldsymbol{a}_i \tag{1.1}$$

which states that it is possible to determine the acceleration $\boldsymbol{a}_i$ of a particle i of mass m_i by knowing the force $\boldsymbol{F}_i$ applied on it. Since the force acting on each atom of a macromolecule can be expressed as the first derivative of the potential energy V of the system with respect to the atom position $\boldsymbol{r}_i$, and the acceleration corresponds to the second derivative of the atomic position with respect to the time, equation (1.1) can be rewritten in its differential form as:

$$-\frac{dV}{d\boldsymbol{r}_i} = m_i . \frac{d^2\boldsymbol{r}_i}{dt^2} \tag{1.2}$$

For atoms with positions $\boldsymbol{r}_i$ at a specific time t, the position a small time step Δt later is given by a second order Taylor expansion.

$$\boldsymbol{r}_i(t + \Delta t) = \boldsymbol{r}_i(t) + \boldsymbol{v}_i(t).\Delta t + \frac{1}{2}\boldsymbol{a}_i(t).\Delta t^2 \tag{1.3}$$

This enables the atomic positions to be calculated numerically after a small time step knowing the position, velocity and acceleration at time t, the acceleration being obtained with Newton's second law. Therefore, starting from an initial atomic model prepared from X-ray crystallography or NMR data, and from initial velocities obtained from a Boltzmann distribution at a given temperature, a trajectory can be calculated by derivation of the potential energy function with respect to the three-dimensional position of each atom.

However, equation (1.3) is, in fact, a truncated version of the Taylor expansion, which includes higher order terms. These terms cannot be simply ignored, as it would lead to important errors. One computationally efficient way to overcome this is to use the Verlet algorithm or one of its variants. Presently the velocity Verlet algorithm is used in most of the MD simulation packages owing to its improvement over the classical Verlet algorithm and its leap-frog variant. Indeed, the Verlet algorithm uses no explicit velocities and lack of precision, whereas in the leapfrog variant, velocities are explicitly calculated but they are not calculated at the same time as the positions. In contrast, the velocity Verlet algorithm does not compromise on precision and velocities, accelerations and positions are calculated at the same time t by relying on equation (1.3) and also on equation (1.4), which gives the velocities $\boldsymbol{v}_i$ at a specific time t plus a small time step Δt later.

$$\boldsymbol{v}_i(t + \Delta t) = \boldsymbol{v}_i(t) + \frac{1}{2}[\,\boldsymbol{a}_i(t) + \boldsymbol{a}_i(t + \Delta t)\,].\Delta t^2 \qquad (1.4)$$

The only missing piece is the potential energy of the system to retrieve the acceleration. Even though potential energy can be calculated using quantum mechanics in order to take into account electronic effects, this is only possible for small molecules due to the computational resources required. For proteins, the potential energy can only be calculated by empirical energy functions based on classical mechanics. Both the energy arising from bonded and from non-bonded interactions should be taken into account as shown in the simplified, typical force field described in Figure 1.7 [80]. Bonded interactions, bond stretching and angle bending are modelled using harmonic potentials, while the steric barriers (as well as electronic effects depending on the force field) between atoms separated by three bonds are modelled by a torsion angle potential. Non-bonded interactions are composed of van der Waals interactions, modelled by the Lennard-Jones 6-12 potential, and of electrostatic interactions represented by a Coulomb potential. Parameters used in the potential energy function are determined using both quantum mechanical calculations on small systems and experimental data. The potential energy equation and its parameters are collectively referred to as a force field. Several force fields are available and each of them is based on slightly different equations and parameters, the most common include AMBER [81], CHARMM [82] and GROMOS [83].

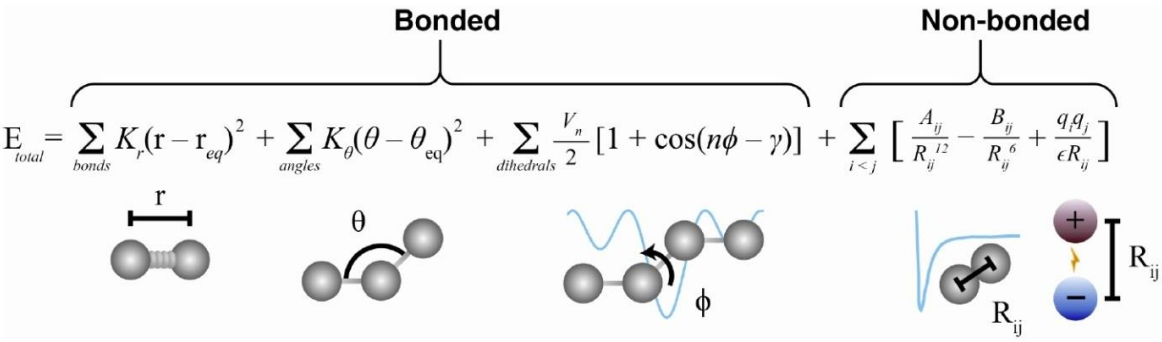


Figure 1.7: Typical molecular mechanics force field. Example of an equation used to approximate the potential energy function of a protein (from reference [80]. Chemical bonds and atomic angles are modelled using harmonic potentials, whereas dihedral angles are modelled using a sinusoidal function. Van der Waals interactions are modelled using the Lennard-Jones potential and electrostatic interactions using Coulomb's law. (© Durrant and McGammon; licensee BioMed Central Ltd. 2011).

The trajectory of each atom can then be solved numerically using a MD software packaged such as AMBER [84], CHARMM [85], GROMACS [86,87], GROMOS [88], or NAMD [89], all of which are specifically designed to run MD simulations on supercomputer clusters. In addition, in order to integrate the equations of motion, MD codes allow the microscopic data (atom positions, velocities, etc.) obtained from the simulation to be linked to macroscopic observables (energy, temperature, pressure, etc.), using statistical mechanics. At this point, it is important to note that MD simulations calculate time averages while the macroscopic observables are assumed to be an ensemble average, therefore, to link the results of a MD simulation to macroscopic data, one should rely on the ergodic hypothesis which states that ensemble average equals time average. To satisfy this equality, MD simulations should be sufficiently long to sample enough representative conformations.

Using the best-in-class supercomputers, classical all-atom MD studies can hardly reach timescales of up to a few microseconds for protein systems. They are therefore particularly adapted for sampling the free energy landscapes corresponding to tier-1 and tier-2 dynamics. However, since the simulated timescale should be several times longer than the slowest event of interest, tier-0 dynamics remain mainly inaccessible with this technique. Consequently, when this is required, classical all-atom MD must be modified to access rare events. In order to access long timescales with reasonable computational resources, several methods have been

developed which are usually based on an enhanced sampling of the free energy landscape or on the clustering of short trajectories by numerical methods based on probability theory.

1.4.1.2 Enhanced sampling techniques based on MD

Enhanced sampling techniques consist of modifying the classical all-atom MD procedure in order to speed-up the sampling of the free energy space. As these methods introduce bias and thus alter the kinetic or thermodynamic results obtained from the simulations, it is often required, for example, to combine different techniques or to perform a reweighting procedure.

1.4.1.2.1 Targeted MD

In targeted molecular dynamics (TMD), both an initial and a final state have to be known, and the system will be driven toward the targeted conformation by the application of a bias force which is a function of the difference between the initial and final coordinates (represented by the RMSD value) is applied to each atom as shown in (1.5):

$$V_{TMD} = \frac{k}{2N} [RMSD(t) - RMSD^*(t)]^2 \quad (1.5)$$

where V_{TMD} is the bias potential, $RMSD(t)$ is the RMSD from the target at time t, $RMSD^*(t)$ corresponds to the linear evolution of RMSD between the initial and the target structure, k is the spring constant and N the number of atoms affected on which the TMD force is applied.

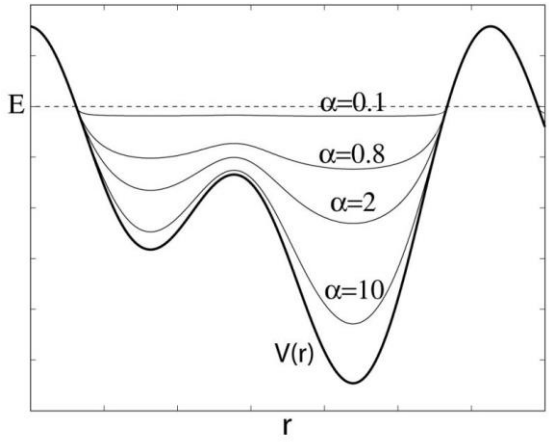
This RMSD-dependent force is greater for large-scale global motions that will decrease the RMSD more quickly than small-scale localised changes and therefore large motions will occur first and small ones will occur later while more realistically a combination of both occur

over time. Therefore, the computed path depends on the direction (initial to final or final to initial) and does not truly describe the conformational transition. Thus extensive additional work is required to better describe the system and eliminate the bias [90,91].

1.4.1.2.2 Accelerated MD

The accelerated molecular dynamics [92] (aMD) method applies a bias potential if the potential energy falls below a threshold energy defined by the user; it raises the energy landscape of deep wells while leaving the rest of the landscape unaffected. As a result, the barrier separating adjacent energy basins is lowered, allowing the system to escape low energy wells more easily and to sample a larger part of the conformational space when compared with classical MD.

The challenge of aMD is to find the appropriate acceleration parameters for the studied system, i.e. the acceleration factor α which determines the shape of the modified potential and the threshold energy E which controls the portion of the potential surface affected by the boost (Figure 1.8). The choice of these parameters is critical since, if not chosen wisely, only little enhancement of the sampling could be observed or instead the system could explore regions that are not of interest and lead for example to the partial unfolding of the protein [93].



$$\Delta V(\mathbf{r}) = \begin{cases} 0 & V(\mathbf{r}) \geq E \\ \frac{(E-V(\mathbf{r}))^2}{\alpha+E-V(\mathbf{r})} & V(\mathbf{r}) < E. \end{cases}$$

Figure 1.8. Schematic representation of accelerated MD. In the aMD method, when the original potential (thick line) falls below the threshold energy E (dashed line), a boost potential ΔV(r) is added. The shape of the modified energy profile (thin lines) is determined by the acceleration factor α (source: NAMD manual http://www.ks.uiuc.edu).

For example, aMD has been used by Gasper *et al.* to study the allosteric regulation of thrombin upon association with thrombomodulin [94]. In this work two different boost potentials were used, one for the dihedral energy and the other, weaker, for the total energy of the system. The parameters $E(dih)$ and $\alpha(dih)$ for the dihedral boost and $E(tot)$ and $\alpha(tot)$ for the boost on the total potential were chosen so:

$$E(dih) - < Vo(dih) >= x \times \ Number\ of\ residues \quad (1.6)$$

$$\alpha(dih) = \frac{1}{5} \times \ E(dih) - < Vo(dih) > \quad (1.7)$$

$$\alpha(tot) = E(tot) - < Vo(tot) > = 0.16\ kcal.mol^{-1} \times \ Number\ of\ atoms \quad (1.8)$$

where $< Vo(dih) >$ and $< Vo(tot) >$ represent the average dihedral and total potential energies respectively, and were obtained after a short classical MD simulation. Two levels of acceleration were chosen for the dihedral boost, a moderate and an aggressive one, where x was set to 2 and 4 kcal.mol^{-1} respectively. These two levels of acceleration were used to access the motions occurring at the hundreds of nanosecond (moderate acceleration) and tens of microseconds (aggressive acceleration) timescales and were used in conjunction to classical MD to study fast and slow motions relevant to the allosteric regulation of thrombin.

1.4.1.2.3 Metadynamics

In metadynamics [95], a boost potential is added to the energy well corresponding to the present location of the system. This history-dependant bias energy takes the form of small potential hills that flatten the already visited energy landscape and prevent the system from revisiting it. So effectively, metadynamics "fill" the energy basins as they are visited, but this can only be performed on a limited number of conformational coordinates, otherwise it would be too expensive computationally and difficult to interpret. The challenge is to identify the conformational coordinates (collective variables) that describe correctly the motions of interest of the protein and require therefore *a priori* knowledge. Metadynamics also suffers from convergence problem, but this is in a large part solved by using its variant, well-tempered metadynamics [96]. One of the most interesting features of metadynamics is that the free energy changes along the selected coordinates can easily be retrieved as it corresponds to the negative mirror image of the energy bias [97], and recently a method has been developed to retrieve transition rates from metadynamics simulations [98,99].

A variant of metadynamics, well-tempered ensemble [100], has recently been used to study the allosteric regulation of the KIX domain of the CREB-binding protein that takes place on the millisecond timescale with limited computational resources [101]. The authors were able to compare their results to NMR experiments and to characterize structurally and dynamically an excited state relevant to allostery, which could not be characterized by NMR owing to its short lifetime.

1.4.1.2.4 Replica-exchange molecular dynamics

During temperature replica-exchange MD (T-REMD) [102] simulations, N replicas (series of simultaneous non-interacting simulations) are generated, over a range of user-defined temperatures. At a given exchange rate, adjacent replicas are allowed to exchange their conformation, which induces large thermal fluctuations, which in turn, assist replicas in escaping low energy basins and optimize the free energy landscape sampling. Although kinetics information cannot be extracted directly from T-REMD experiments due to the temperature swaps between MD trajectories, recent developments allow the generation of pseudo-kinetics that are sufficient for the kinetic clustering of the system [103]. In order to perform well, T-REMD requires a good exchange acceptance ratio between replicas (typically > 20%) which corresponds to the number of successful attempts to switch between replicates on the total number of attempts to switch, and the larger the system the higher the number of replicas required to maintain a ratio above 20%. Therefore, T-REMD is essentially limited to relatively small systems. To overcome this limitation, other replica methods that are not based on temperature changes have been reported in the literature [104], such as Hamiltonian replica-exchange (H-REMD).

A H-REMD method was recently used to study the disorder-to-order transition of an intrinsically disordered loop of sortase A [105]. This approach enabled to sample a significantly larger region of the conformational space than a classical MD simulation of the same length and allowed to decipher the allosteric regulation mechanism of the protein.

1.4.1.3 Coupling Markov state models with conventional MD

The main strategy developed within the past few years, mainly in the field of protein folding [106], to access long timescales is to run very short simulations, each sampling different parts of the free energy landscape and then to reconstruct the whole process using Markov state models (MSM) [103,107,108]. MSM is mainly based on a kinetic clustering of discrete conformations obtained from several short MD runs. Conformations that can interconvert rapidly, thus belonging to the same energy basin, will be grouped together into the same state. A transition probability matrix is constructed and gives the probability of transitioning from a state to another during a time Δt chosen to make sure that the system is memory-less (or Markovian – i.e. the probability of transition does not depend on how the system arrived in its initial state). From this model the pathways and kinetic rates between different energy basins can be determined.

Different techniques have been used to run the short MD simulations that indeed sample adequately the energy landscape, for example the folding@home project relies on running several MD simulations in parallel, as soon as one experiences a conformational change identified by monitoring heat capacity spikes associated with the crossing of free energy barriers, all the simulations are reset and restarted from the new coordinates obtained following the transition in order to explore the new basin identified. So in practice, instead of running

one extra-long trajectory using N processors, N very short trajectories are run on a single processor [109]. Another approach consists in using MSM-driven adaptive sampling [110]: MSM models are generated on the fly during the MD simulations and enables to identify regions of the conformational space that are underexplored, permitting simulations specifically to be started to explore these areas. Yet another method recently employed is based on performing a morphing between two end structures of interest and to start simulations from coordinates along the pathway, in order to sample primarily the region of the landscape relevant to the function of interest [111].

The Amaro group reported the study of allostery of a prototypical signalling domain based on an ensemble of MD simulation combined with MSM [112]. Their results show that both the apo and allosteric effector-bound systems are able to sample the active and inactive states, but the presence of allosteric effector favours the active state while its absence favours the inactive one, indicative of a conformational mechanism of allostery. In addition, the authors showed that although the presence of the allosteric effector permits the sampling of the inactive state, it slows down significantly the active-to-inactive transition. The authors were able to draw these conclusions because the combination of adaptive sampling and MSM permitted a thorough sampling of the energy landscape.

1.4.1.4 Revealing allosteric communication pathways by analysing MD trajectories

Exploring the energy landscape is a first step towards the understanding of allosteric regulation mechanisms in proteins, however, except in the case of obvious conformational changes, it is not always sufficient to understand the communication of the allosteric information from the allosteric binding site to the catalytic site. One of the great advantages of

MD simulations is that it gives an atomic level resolution of protein dynamics, however, it means that a tremendous amount of information is generated but only a fraction is interesting from a functional point of view. Therefore, powerful analysis techniques [70] are required in order to discriminate between motions relevant to allosteric regulation and the rest.

1.4.1.4.1 Correlated motions between residues

One of the most popular approaches consists in studying correlated motions between residues. The idea being to identify the residues that move in a concerted manner along the pathway between allosteric and active sites. Traditionally, correlated motions are identified using the Pearson coefficient, unfortunately it means that only linear correlations will be accounted for and that the potentially important more complex, non-linear correlated motions will not be taken into account. In order to study non-linear correlated motions, mutual information, which is based on information theory [113], can be used as a measure of correlation. It has been used to identify alpha carbon correlated motions [114] and more recently, with allostery specifically in mind, correlated motions between dihedral angles (including those of the side chains) [115].

1.4.1.4.2 Dynamical network analysis

Correlated motions give an account of the correlated inter-residue movements, but it might be complex to interpret as correlated residues might be very distant in space from one another, making it impossible to identify an allosteric pathway by solely relying on this approach. This was solved by Sethi and co-workers who resorted to graph theory to identify allosteric pathways [116]. In this approach, referred to as dynamical network analysis, each

residue corresponds to a node of the network (or graph) and non-adjacent residues that are in contact (based on an inter-residue distance falling below a specific cut-off distance) represent the edges of the network, which are weighted as a function of the degree of correlation between the motions of these residues. The obtained network can then be analysed in terms of optimal and suboptimal paths to identify allosteric communication pathways (Figure 1.9). They can also be decomposed in communities of highly interconnected nodes to coarse grain the results of the network analysis. Initially applied to alpha carbons [116], this approach was soon advantageously extended to the centre of mass of each residue [117] and the approach was partially modified to analyse the ensemble of suboptimal paths rather than the optimal one in order to identify hub residues that are involved in a large number of suboptimal paths, and are therefore key to allosteric signal propagation [118].

This approach was used by the McCammon group in their study of allosteric regulation in thrombin [94]. Instead of using the Pearson coefficient, the authors relied on mutual information to identify correlated motions between residues. Using this approach, they were able to identify two allosteric pathways, rationalising the results from H/D exchange experiments. Dynamic community network analysis was also used more recently to provide a mechanistic explanation of the how the Ras protein is affected by an oncologic allosteric point mutation [119].

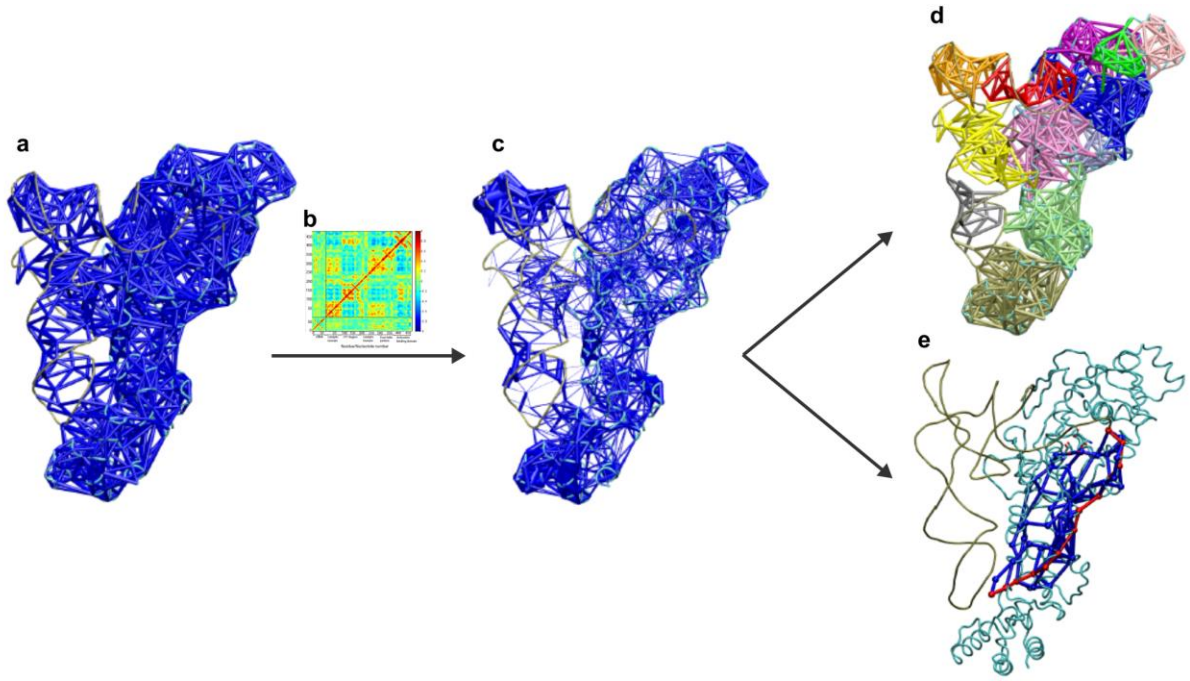


Figure 1.9. Dynamic network analysis. Each node of corresponds to a residue (or nucleotide) and the edges between nodes indicates residues in contact within a given cut-off distance (a). The correlations data between residues (b) is used to weight the network (c). The resulting network can be coarse grained in communities (d) which have stronger connections with the other nodes belonging to the same community than with the nodes belonging to other communities. In addition, optimal (red) and sub-optimal (blue) communication paths between two given residues can be identified. (images: tutorial on "Dynamical network analysis" http://www.ks.uiuc.edu).

1.4.1.4.3 Other approaches relying on networks of correlated motions

Two approaches that are related yet different to the dynamical network analysis have been developed to identify communication pathways in proteins. The first one relies principally on information theory [120], first local correlated motions are determined using mutual information, and the network of correlated residues and shortest path are calculated as per the dynamical network analysis. In addition, the correlation between local and global motions are

also included in the analysis using once again mutual information, with the global motions being extracted from principal component analysis.

In the second approach, named MONETA [121], the network is composed of clusters of residues, each cluster presenting residues that have highly correlated motions but presenting weakly correlated motions with the residues found in other clusters. These clusters are identified using local feature analysis (a method used in image processing) from the principal component analysis. These clusters are then connected via chains of communicative residues obtained by calculating the inter-residue distance variance.

1.4.1.4.4 Frustration

In another approach, the concept of frustration in folded proteins [122], has been applied to allostery [123]. This approach seeks to identify regions of high frustration, which arise when the contact between two residues is not energetically optimised. The rationale behind this is that upon binding of an allosteric ligand, the level of frustration in these regions will decrease, enabling the identification of residues involved in allosteric regulation. The authors used this method on a number of allosteric proteins known to deploy large-scale conformational changes upon allosteric action and found that highly frustrated regions tend to be the regions that reconfigure locally upon allosteric regulation.

1.4.1.4.5 Force directed analysis

Force directed analysis (FDA) has been developed as a method to visualise strain propagation within a protein [124]. More specifically FDA works by calculating pairwise

atomic forces within the structure and seeks to identify residues that experience large changes in forces upon binding of an allosteric ligand. This enables the identification of signal propagation even in the absence of major atomic displacements.

This approach has been used to identify allosteric communication pathways within various proteins [125-127], in particular in the case of the allosteric activation of the methionine repressor which does not experience any major conformational changes, the authors were able to identify using this method a number of residues which had been shown to be essential in the allosteric communication via mutagenesis experiments [128].

1.4.1.4.6 Other approaches

Various other methods have been used to study allostery from MD simulations, such as the study of the dynamic correlation of residue-residue contact events [129], identification of directly coupled side chains [130] or N-body information theory analysis [131]

1.4.1.5 Alternative methods

A number of techniques involving modified MD procedures have been developed to directly identify allosteric pathway within proteins, for example, using anisotropic thermal diffusion [132] or rigid residue scan [133]. But MD is not the only method that can study allostery via illumination of protein motions, and a number of approaches that require significantly less computational resources have been reported.

Using a Gō model, Okazaki and Takada showed [134] that binding of small molecules tends to follow a conformational selection model, whereas protein-macromolecule binding is more likely to proceed via induced fit. The conclusions of this work have been in part contradicted by the recent study of a protein-protein binding using NMR [135], which demonstrated that a conformational selection mechanism was at play. Nevertheless, the simplicity of the model made amenable the modelling of several binding and unbinding events in the presence of absence of conformational changes, within a limited computational time.

The group of Andrej Sali has developed a hybrid structure-based / Gō model defining simplified 2-wells energy landscapes based on the crystal structure of the bound and unbound form of an allosteric protein [136]. The energy landscape can then be explored using MD starting from different points along this landscape and effective sampling can bridge the gap between microscopic motions and macroscopic allosteric mechanisms.

Elastic network models (ENM), which are a coarse-grained model for normal mode analysis, have been used with some success to describe allostery. As they enable to assess with good accuracy the large amplitude, low-frequency modes, ENM models are particularly well suited to study dynamic allostery [137-139].

1.4.2 Exploring the catalytic energy landscape

1.4.2.1 Modelling enzymatic reactions to understand allostery

In the case of an enzyme, allosteric regulation usually implies that action of an allosteric effector will affect the catalytic activity of the enzyme, in which case an in-depth understanding of the reaction taking place within the enzyme can be extremely valuable to understand how

conformational or dynamical changes at the active site can affect the energy profile of the reaction. As molecular mechanics (MM) is based on classical mechanics, this approach cannot be used to model bond formation or cleavage and therefore in order to model the electrons of the system, one needs to resort to quantum mechanics (QM).

QM calculations are particularly intensive from a computational point of view and the required computational power increases significantly with the size of the system. It is therefore typically not feasible to model a complete protein (and surrounding solvent) with QM. However, changes occurring during catalysis are usually confined to a small region of the enzyme with only a few residues directly involved. Therefore, a very powerful approach to study enzymatic reactions is to use a hybrid QM/MM scheme in which the active site and substrates are modelled using a QM level of theory, whereas the rest of the solvated enzyme is modelled at the MM level of theory. This approach enables to model in a computationally efficient manner the reactions taking place within the active site of an enzyme.

1.4.2.1 Brief overview of the foundations of quantum chemistry

The theoretical underpinning and computational implementation of quantum chemistry have been detailed in several classical textbooks [76,140,141] and here only some of the key concepts relevant to the work presented in this thesis are introduced.

Quantum chemistry is typically based on the time-independent Schrödinger equation, which requires a number of approximations to be applied in order to find a solution. The first one, the Born-Oppenheimer approximation, consists in considering the nuclei as fixed in space.

Indeed, due to the difference in mass between nuclei and electrons, the electrons will move faster and thus their motion can be considered as being instantaneously equilibrated at each nuclear displacement. This approximation helps to greatly simplify the Schrödinger equation, but is not sufficient. The next approximation, the orbital approximation, consists in considering the wave function of a many-electron system as the product of one-electron wave functions called orbitals. In the third approximation, each orbital can be described by a linear combination of a set of atomic orbitals (LCAO) that form a basis set.

These approximations form the basis of the Hartree-Fock (HF) method. HF, however, does not take into account electron correlation, i.e. the tendency of electrons to avoid each other, and this omission can lead to significant errors. To overcome this problem, other *ab initio* methods, such as the 2nd order Møller-Plesset (MP2) method, take into account electron correlation at the cost of an increased computational burden. Density functional theory methods (DFT) have been developed with the objective of approaching the accuracy of the post-HF *ab initio* methods at a lower computational cost. These methods rely on the electron density distribution instead of the wave function in order to calculate the (ground state) energy. However, the functional linking electron density and energy is not known and the different DFT methods seek to approximate it. Several functionals have been developed, among which B3LYP is probably the most famous. Semi-empirical methods (such as MNDO, AM1, PM3, PM6, DFTB etc.), are the least computationally demanding methods. MNDO-types methods are based on a simplification of the HF method whereas DFTB is based on DFT. Semi-empirical methods are typically the least accurate and rely on experimental correction to improve accuracy. However, utilisation of a semi-empirical method on a reaction for which the method has been correctly parametrised can lead to results that approach the reliability of high-level calculations for the fraction of the computational burden.

1.4.2.2 Hybrid quantum mechanical/molecular mechanical simulations

QM/MM methods have been presented in great detail in various excellent reviews [142-147] and I focus here on the approaches used in this thesis. Hybrid QM/MM approaches for modelling enzyme reactions are based on using a QM method to describe the substrates as well as the residues involved in the reaction and an MM force field to describe the rest of the system. The size of the QM region, i.e. which atoms are treated by QM, is a trade-off between accuracy and computational cost: the more atoms added the longer the calculations, but the better the accuracy might be. Usually, the substrates, the residues participating to the reactions (in most of the cases their side chain only) as well as other molecules that might be directly involved in the reaction (metal ions, water molecules, inorganic ions, etc.) are included in the QM region.

The total energy of the system can be expressed as follows (in the case of the additive QM/MM approach relevant to this thesis):

$$E_{Total} = E_{QM} + E_{MM} + E_{QM/MM} + E_{Boundary} \quad (1.9)$$

where E_{QM} corresponds to the energy of the QM region, E_{MM} to the energy of the MM region, $E_{Boundary}$ arises from the finite number of atoms in the simulation, and $E_{QM/MM}$ describes the interactions between the QM and MM regions. This term accounts for the electrostatic "embedding" of the QM region within the MM region, allowing the QM region to be affected by the charges of the atoms in the MM region. It also includes van der Waals interactions between QM and MM atoms as well as bounding terms arising when the separation between QM and MM region "cut" a covalent bond (e.g. when only the side chain of a residue is included in the QM region). In this case, the most common approach consists in creating a

fictitious link atom (usually a hydrogen atom), which is added along the covalent bond that crosses the QM-MM boundary, so the valence of the QM atom is satisfied. It is important to note that one should avoid "cutting" polar covalent bonds and, when possible, a good choice consists in cutting a C-C bond distant from any functional groups or from the backbone atoms.

There are two main types of QM/MM approaches to model reactions. The first one, adiabatic mapping, essentially relies on performing energy minimization at different values of a reaction coordinate to reconstruct the potential energy surface. The advantage of this approach is that high-level QM methods can be used (*ab initio*, DFT). The main drawbacks are that the energy minimization can lead to local minima (and thus conformational changes during the reaction cannot be correctly modelled) and the calculations lead to a potential energy profile and not a free energy profile, although the free energy profile can, for example, be estimated by performing frequency calculations in order to include an approximated entropy contribution. The second approach is similar but consists in running MD simulations, more specifically umbrella sampling along the chosen reaction coordinate (QM/MM-MD approach). This has the advantage that it leads to a free energy profile after statistical reweighting of the umbrella sampling trajectories, using for example, the Weighted Histogram Analysis Method (WHAM), and to a more accurate exploration of the phase space. However, this method requires a large number of calculations of the energy (one per MD step) so only less accurate, semi-empirical methods (e.g. AM1, PM6, SCC-DFTB, etc.) can be used.

As conformational and catalytic landscapes are linked and different conformations can lead to different energy profiles of reactions [148-150], it is important several QM/MM calculations starting from slightly different conformations are performed (e.g. obtained from

MD simulations) to account for the fact that, in reality, the reactions do not start from a single structure but instead from an ensemble of structures.

QM/MM has, for example, been used to study the Fe-CO bond dissociation energy in myoglobin [151] and the authors showed that the dissociation was more facile in the T state than the R state with an energy lowered by 2.6 kcal/mol, essentially because of steric effects, accounting for the existence of two different allosteric states. More recently, it has been shown using QM/MM that allosteric phosphorylation of the insulin receptor kinase enhances the catalytic rate of the enzyme by affecting the dynamics of the protein, which in turn significantly changes the free energy landscape of the first step of the reaction, thereby directly showing how dynamic allostery can impact enzyme catalysis [152].

1.4.3 Exploring the evolutionary landscape

The processes by which allosteric regulation has emerged in unregulated enzymes can be analysed from an evolutionary perspective. Evolution has led to the development and fine-tuning of various allosteric regulation mechanisms, from large domain motions to subtle changes in protein dynamics. Yet despite its importance, little is known about how allosteric regulation has been acquired and has evolved in proteins [153].

Recent findings provide some clues on how allostery can evolve. For example, from a structural point of view, it is thought that three mechanisms can occur: expansion of flexible or unused parts of the protein (e.g. N- or C- terminal domains, loops); formation of homo- or hetero-protomers which give rise to new potential binding sites and communication pathways; gene fusion to create multi-domain proteins [154]. In addition, it appears that latent allosteric

diversity in proteins can be exploited by novel regulators, initiating a selection process producing divergent allosteric responses [155]. It has also been shown that allosteric site residues tend to be less conserved than catalytic site residues suggesting they evolve along different pathways [156].

A number of examples are found in the literature in which evolutionary analyses, either at a residue level via co-evolution, or at a protein level via phylogenetic analysis, has provided valuable insights into the allosteric regulation mechanisms of proteins.

1.4.3.1 Co-evolved residues

One of the most popular approaches that link evolution to allostery is called statistical coupling analysis (SCA) and was originally developed to identify energetically coupled residues [157] but was soon extended to the identification of allosteric regulation pathways by monitoring residues that have co-evolved [158]. Co-evolution at the residue level, or correlated mutation, primarily arise when two residues are entangled and the relationship between these residues needs to be preserved to maintain function [159]. As a simple illustrative example, let us consider a Lys residue, which upon binding of an allosteric ligand forms a salt bridge with a Glu residue, essential to communicate the allosteric information. If the Lys happens to be mutated into a Glu, then the Glu residue may in turn, be mutated into a Lys so the key regulatory interaction is conserved. SCA has received considerable interest in its original formulation and has been later updated to define the concept of protein sector, corresponding to networks of coevolving adjacent residues that can provide a framework for the evolution of allosteric sites in proteins [160,161]. SCA has also been used in conjunction to MD to identify residues correlated both from a dynamical and evolutionary point of view [162].

In a recent example of application [163], SCA performed on a family of cysteine peptidases led to the identification of protein sectors and the surface residues belonging to these sectors were used to identify potential allosteric binding sites in cathepsin K which belongs to this peptidase family. Computational docking permitted the identification of a novel inhibitor whose inhibitory effect was confirmed experimentally.

1.4.3.2 Phylogenetic analysis

Apart from SCA, more classical bioinformatics approaches such as phylogenetic analysis or ancestral sequence reconstruction can be used advantageously to shed light on allosteric regulation mechanisms and evolution. For example, using phylogenetic analysis, ancestral sequence reconstruction and structural biology, Thornton and co-workers [164] studied how a steroid hormone receptor, which is allosterically regulated in vertebrates, evolved into an unregulated one in molluscs. They were able to identify the residues mutations causing the protein to remain constantly switched on in molluscs, effectively deciphering part of the allosteric regulation mechanism at the molecular level.

1.5 Aims of this thesis

The overall objective of this thesis is to provide new insights into the understanding of allostery by studying different allosteric regulation mechanisms using a variety of computational methods. The metabolic enzyme 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase (DAH7PS), which catalyses the first step of the shikimate pathway, is used as a model

enzyme for allostery due to the unprecedented variety of allosteric regulation mechanisms this protein can deploy, from large domain movements to dynamic allostery, representing an ideal playground to study allostery as discussed in Chapter 2.

However, with the notable exception of *M. tuberculosis* DAH7PS, which has been studied by molecular dynamics (MD) simulations [165], the different allosteric regulation mechanisms are unclear and have only been deduced from static models obtain by X-ray crystallography. Although these models can help envisage possible effects of allosteric inhibition, for example, based on the structure of an initial and a final state as well as on the temperature factors displayed by different parts of the protein in the crystal structure, they lack information about the actual dynamic process occurring within the protein and which is of upmost importance for allosteric signal propagation. For instance, MD simulations of *M. tuberculosis* DAH7PS, which gave access to the protein dynamics, have played a key role in deciphering the complex synergistic regulation of type II DAH7PS enzymes.[165]

In Chapter 2, The DAH7PS family and its metabolic importance are presented, and its relationship to other phosphoenolpyruvate aldolases is explored. Then, using a variety of sequence and phylogenetic analysis tools, the phylogenetic relationship between members of the DAH7PS family is revisited in light of the DAH7PS sequences and structures available today, and the evolution of allosteric regulation in this enzyme family is discussed.

In Chapter 3, a new method to identify allosteric communication pathways in protein relying on dynamic allosteric regulation is presented. This approach, which is based on the

monitoring of calculated pK_a variations of ionisable residues in proteins, is used to decipher the allosteric regulation mechanism of DAH7PS from pathogen *Neisseria meningitidis*.

In Chapter 4, Using MD simulations, the allosteric regulation of the DAH7PS from hyperthermophile *Thermotoga maritima* is brought to light, and reveals hidden complexity in the seemingly cumbersome mechanism that involves large amplitude domain movements.

In Chapter 5, DAH7PS from thermophilic soil bacterium *Geobacillus sp.* is studied using X-ray crystallography, homology modelling and small-angle X-ray scattering in the presence and absence of an allosteric inhibitor, revealing domain capping as well as the folding of a flexible region of the regulatory domain. The strategy used to explore the allosteric transition, based on MD simulations, currently in progress are also presented.

In Chapter 6, different attempts to model the first step of the reaction catalysed by DAH7PS are described and recommendations for future work toward the complete modelling of the reaction is provided.

Chapter 2 – Revisiting the phylogenetic relationship between members of the DAH7PS family

2.1 Introduction

2.1.1 Aromatic amino acids and the shikimate pathway

The three proteinogenic aromatic amino acids phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp) are essential to all living cells. Animals are not capable of making Phe and Trp, which therefore must be acquired through diet (Tyr is a conditionally essential amino acid as it can usually be biosynthesised in animals from Phe via an enzymatic hydroxylation step) [166]. Microorganisms and plants, however, do not rely on diet to acquire aromatic amino acids, but instead utilise the shikimate pathway to produce the essential intermediates for the biosynthesis of the three aromatic amino acids as well as a variety of important compounds such as vitamins B_9 and K_1 [167].

The shikimate pathway consists of seven enzymatic steps which ultimately convert erythrose 4-phosphate (E4P) and phosphoenolpyruvate (PEP) into chorismate, the key intermediate in the biosynthesis of the three aromatic amino acids (Figure 2.1) [167]. As this

pathway is only present in bacteria, fungi, plants and in some apicomplexan parasites but not in mammals [167-169], it represents a key target for the development of novel antibiotics [170,171], antifungal and antiparasitic agents [172], but also of herbicides [173]. The potential for drugs and pesticides targeting the shikimate pathway is particularly well illustrated by glyphosate [174], marketed in 1974 by Monsanto under the tradename Roundup®, which is perhaps the most commonly used herbicide worldwide. Glyphosate targets 5-enolpyruvyl-shikimate-3-phosphate synthase which catalyses the 6th step of the shikimate pathway and owes much of its success to its high potency on an extremely wide range of plants while remaining one of the least toxic herbicides to animals [175].

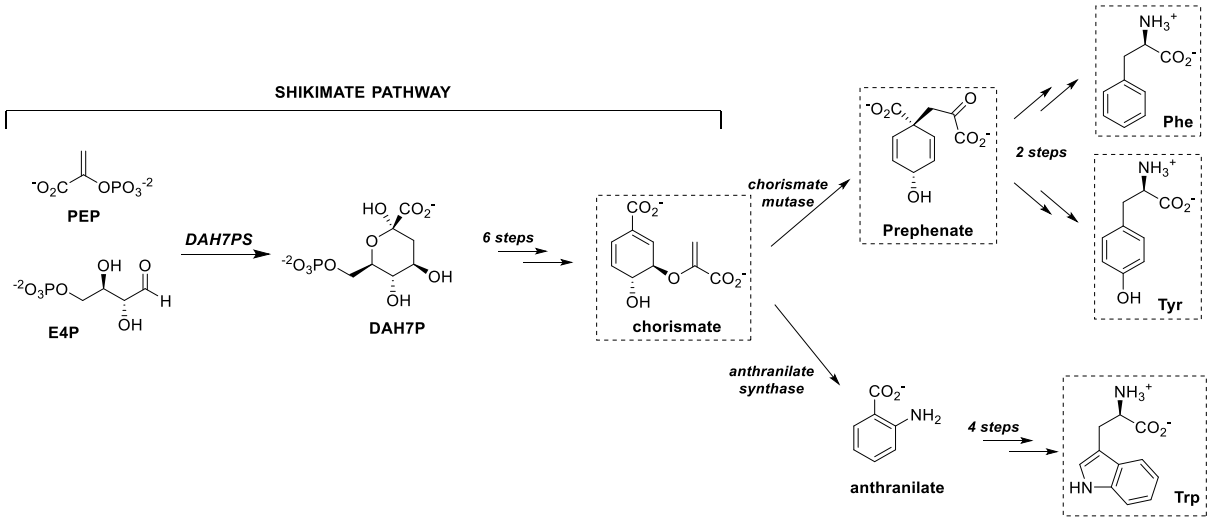


Figure 2.1. Shikimate pathway and aromatic amino acids biosynthesis. E4P and PEP are converted into chorismate in 7 steps, the first reaction being catalysed by DAH7P synthase. Chorismate is then converted into Tyr, Phe and Trp. These end-products, as well as intermediates chorismate and prephenate, have been shown to regulate the shikimate pathway by feedback inhibition of DAH7PS.

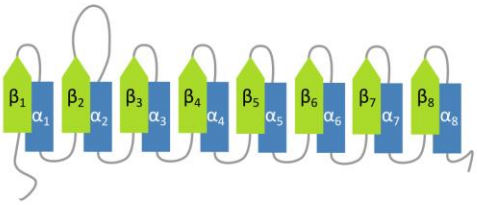
2.1.2 The DAH7PS family

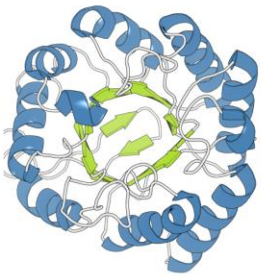
The first committed step of the shikimate pathway is catalysed by 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase (DAH7PS), which catalyses an aldol condensation

reaction between PEP and E4P to yield 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAH7P) and inorganic phosphate. DAH7PS is a metalloenzyme that requires a divalent metal ion to activate the aldehyde functionality of E4P, thereby facilitating the attack of PEP. DAH7PS is a key control point of the shikimate pathway, via feedback inhibition by the aromatic amino acids, chorismate, prephenate but also hydroxyphenylpyruvate, phenylpyruvate and arogenate [167,176]. Due to this pivotal role, an in-depth understanding of the mechanism of allosteric regulation of DAH7PS is of primary importance for the development of new antimicrobials targeting pathogens and parasites that rely on the shikimate pathway.

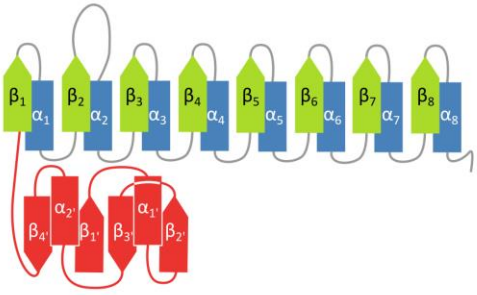
Enzymes of the DAH7PS family are usually classified as type Iα, type Iβ and type II based on sequence analysis. Although all DAH7PS enzymes share a common $(\alpha/\beta)_8$ TIM barrel and an extended $\beta_2\alpha_2$ connecting loop covering the active site, major structural variations exist among the different types of DAH7PS, and even within a single type, which is particularly evident by comparison of the known DAH7PS crystal structures (Figure 2.2). For instance, type Iβ can either be unregulated (absence of regulatory domain) as in *Pyrococcus furiosus* (*Pfu*D) [177] and *Aeropyrum pernix* (*Ape*D) [178], possess a N-terminal ACT regulatory domain which binds Tyr as in *Thermotoga maritima* (*Tma*D) [179,180], a N-terminal chorismate mutase (CM) regulatory domain which binds chorismate and prephenate as in *Listeria monocytogenes* (*Lmo*D) [181] or a or C-terminal CM domain. Type Iα enzymes present a β-hairpin insertion in the $\alpha_5\beta_6$ loop and an N-terminal extension which interacts with the β-hairpin of an adjacent monomer to create an allosteric binding site, as in *Escherichia coli* (*Eco*D) [182] *Saccharomyces cerevisiae* (*Sce*D) [183], *Francisella tularensis* (*Ftu*D) [184] and *Neisseria meningitis* (*Nme*D) [185]. Finally, type II enzymes are equipped with an N-terminal

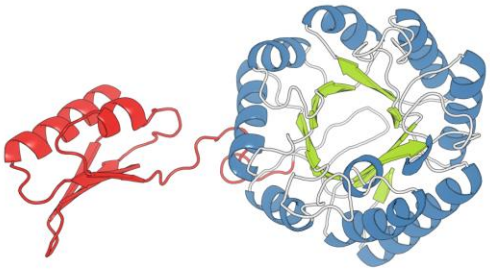
Pyrococcus furiosus - Type Iβ unregulated



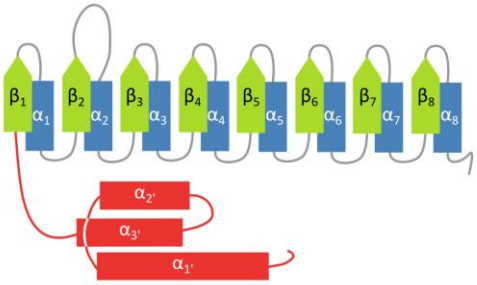


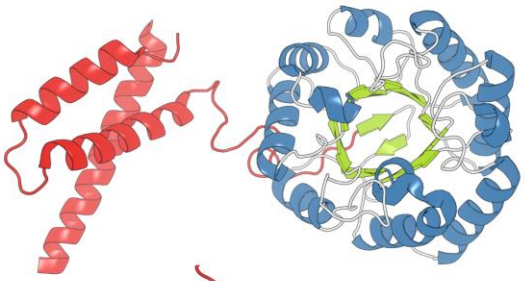
Thermotoga maritima - Type Iβ N-terminal ACT domain



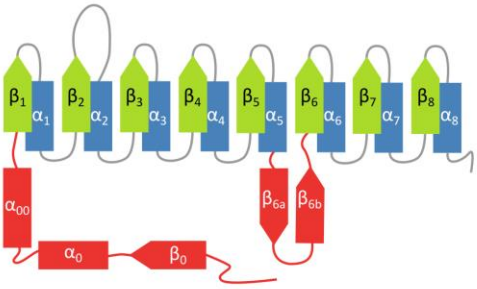


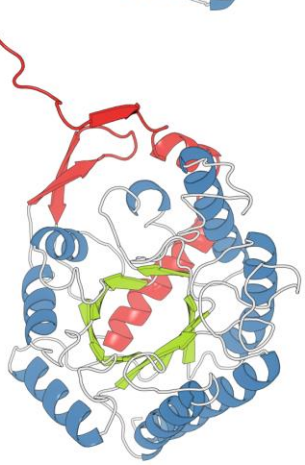
Listeria monocytogenes - Type Iβ N-terminal CM domain



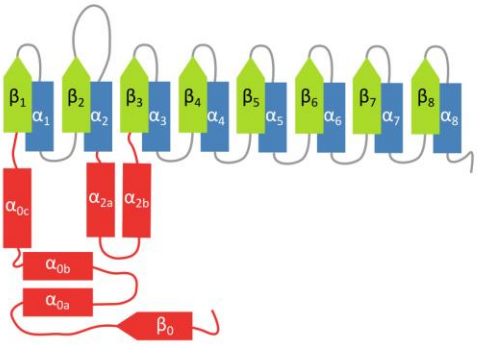


Neisseria meningitis - Type Iα





Mycobacterium tuberculosis - Type II



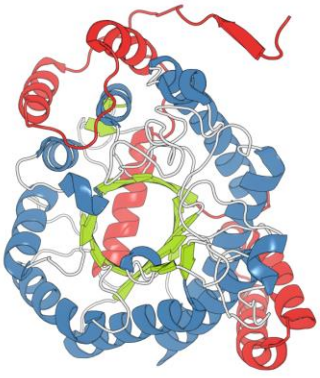


Figure 2.2. Representative structures of DAH7PS. For each subtype, the canonical catalytic barrel is shown in green (β-strands), blue (helices) and grey (loops) whereas allosteric decorations are shown in red. On the left are displayed schematic representations of the secondary structures and on the right the X-ray tertiary structure of the monomeric units are shown.

extension usually coupled to a $\alpha_2\beta_3$ loop insertion, that form the binding site for the aromatic amino acids that act as feedback inhibitors, as in *Mycobacterium tuberculosis* (*Mtu*D) [186] which is synergistically inhibited by a combination of Trp and Phe further increased by the presence of Tyr [187] and is able to recruit CM for a more complex inter-enzyme regulation [188,189]. All the wild-type DAH7PS enzymes for which the structure was solved have been shown to be tetrameric. As these structural insertions or additions are responsible for the binding of allosteric inhibitors and the transmission of the allosteric signal, enzymes of the DAH7PS family exhibit a broad variety of regulatory mechanisms, ranging from drastic conformational changes to subtle changes in protein dynamics.

Various organisms express more than one DAH7PS, it is especially common to find different type Iα isozymes within a given organism, each being allosterically regulated by a different aromatic amino acid. For example, *E. coli* [190] and *Neurospora crassa* [191] express 3 type Iα isozymes, whereas *S. cerevisiae* [192] and *Candida albicans* [193] express 2 type Iα isozymes, one being Tyr regulated and the other Phe-regulated, and *Pseudomonas aeruginosa* expresses two type Iα isozymes, one regulated by Tyr and the other by Trp [194]. Nonetheless, several organisms express only a single type Iα DAH7PS, such as *N. meningitis* [185]. It has been suggested that organisms that rely purely on an endogenous formation of aromatic amino acids use an asymmetric regulation of DAH7PS, i.e. a regulation by a single amino acid, whereas organisms that can utilize aromatic amino acids from exogenous sources have found an effective way to respond to their environment by expressing isoenzymes regulated by each

aromatic amino acids [167]. If this is correct, then organisms that express a single DAH7PS represent better therapeutic targets, as they cannot incorporate amino acids from their environment. Other organisms express both type Iα and type II DAH7PS. It is, for example, the case for *N. crassa* [195], *Stigmatella aurantiaca* [196], *Acanthamoeba castellanii* [197] which also express type II DAH7PS in addition to type Iα. It has been suggested that in this case the type II enzyme is not involved in aromatic amino acid biosynthesis, but instead in the formation of secondary metabolites [198].

Higher plants have also been shown to possess different isozymes, but from type II DAH7PS, as observed in *Nicotiana silvestris* [199], *Arabidopsis thaliana* [200] or *Vigna radiate* [201]. These isozymes, however, appear to play different roles, for example two isozymes have been identified in the plastid, with one continuously expressed and the other expressed in response to infection or wounding [200]. The existence of an isozyme in the cytosol, involved in secondary metabolites has also been postulated, however, no conclusive evidence confirms this hypothesis [168,202]. Plant DAH7PS are not inhibited by aromatic amino acids, but instead can be activated by Trp [168,203].

2.1.3 Relationship between members of the DAH7PS family

The first dichotomy of the DAH7PS family, which distinguished between microbial DAH7PS and plant DAH7PS, arose from the fact that the only microbial (bacterial and fungal) DAH7PS nucleotide sequences known up to the middle of the 1990s belonged to the current type Iα DAH7PS class and thus shared a very low sequence (below 20%) identity with plant DAH7PS (type II) sequences which were also studied at the time [204]. In 1996, Walker and co-workers identified for the first time DAH7PSs in bacteria and fungi that resembled plant

DAH7PS and not bacterial DAH7PS, and proposed a revisited classification: type I DAH7PS have an *E. coli*-like sequence and type II DAH7PS have a plant-like sequence [195].

Two years later, study of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase (KDO8PS) which catalyses the condensation between PEP and D-arabinose 5-phosphate (A5P) revealed that this enzyme is also able to catalyse the condensation between PEP and E4P, suggesting a common origin of the two enzymes [205]. The evolutionary relationship between DAH7PS and KDO8PS was confirmed by sequence analysis and led to the distinction between type Iα and type Iβ DAH7P [205]. Type Iα corresponds to *E. coli*-like DAH7PS whereas type Iβ corresponds to KDO8PS and to a number of DAH7PS resembling *Bacillus subtilis* DAH7PS and having either no regulatory domain or an N-terminal regulatory domain, which can be a CM domain or something else (which was later identified as an ACT domain). It was proposed that the divergence between type Iα and type Iβ occurred prior to the evolution of allosteric regulation within the DAH7PS family, accounting for the existence of unregulated type Iβ and the lack of sensitivity to allosteric control in KDO8PS [205]. Phylogenetic analysis then suggested a microbial origin of type II DAH7PS, proposing that plants acquired DAH7PS via endosymbiosis from Gram-negative bacteria [198].

Birck and Woodard revisited the phylogenetic relationships between DAH7PS on the basis of maximum likelihood (ML)on a set of sequences that included type I and type II DAH7PS and proposed a class I DAH7PS (type Iβ) and class II DAH7PS (type Iα and type II, although the later corresponded to a set of only two structures and diverged significantly from the former) [206], KDO8PS was not included in this tree but instead was studied separately. Their approach was vividly criticized by Jensen, who published an updated tree (based on Neighbour-joining (NJ)), which describes the relationship between type I DAH7PS and

KDO8PS and claimed that type II DAH7PS cannot be included on the premises that the homology between the two types cannot be detected from a sequence point of view and that they should be considered has having evolved independently [207]. Jensen, however, acknowledged that structural determination of type II DAH7PS might change this perspective by providing insights into the divergent evolution between the two types.

This became possible four years later with the structure determination of type II DAH7PS from *M. tuberculosis* (*Mtu*D) by Webby et al. [208], which revealed how surprisingly similar the catalytic scaffold of type I and type II are, indicating a common ancestor between the two types of DAH7PS. In spite of the newfound relationship between type I and II thanks to the structure of *Mtu*D, Roberts and co-workers, who studied the evolution of the shikimate pathway in Eukaryotes, constructed a distinct phylogenetic tree for type I and type II and inferred their evolution separately [209]. Woodard and co-workers refined type I DAH7PS, distinguishing unregulated, N-terminal regulated and C-terminal regulated type Iβ DAH7PS [210], but did not provide any insight onto the type II and type I relationship. Even in a very recent study of the shikimate pathway in prokaryotes fails to reconsider the phylogenetic relationship of the DAH7PS family [211].

It is also important to note that during the first half of the 2000s, the phylogenetic relationship between KDO8PS, which is involved in the biosynthesis of lipopolysaccharides in Gram-negative bacteria, and the NeuB family, which is responsible for the synthesis of *N*-acetylneuraminic acid, became evident [212]. Bacteria utilise NeuB to synthesise *N*-acetylneuraminic acid, which is then found at the end of lipopolysaccharides in order to mimic the cell surfaces of Eukaryotes and prevent the immune response from the host. The reaction proceeds via the aldol condensation of PEP and *N*-acetylmannosamine, however, in contrast

with DAH7PS and KDO8PS in which PEP attacks the re face of the aldehyde moiety of the second substrate, NeuB catalyses the attack of PEP on the *si* face of the aldehyde. Since KDO8PS is related to type Iβ DAH7PS, which was later confirmed by the structure of unregulated *Pfu*D which is extremely similar to KDO8PS structures [177], the phylogenetic relationship between type I DAH7PS, KDO8PS and NeuB was briefly studied [213]. However, to date, no phylogenetic studies of the three DAH7PS types and their relationship to the two other PEP aldolases have been reported.

In the following, I propose to revisit the evolutionary relationship between members of the DAH7PS family in light of the tremendous amount of DAH7PS sequences now available, as well as the numerous structural and biochemical studies of this family conducted since the last phylogenetic trees were proposed. This new analysis details the phylogenetic relationship of type Iα, type Iβ and type II and permits the identification of type Iα and type Iβ subgroups distinct from their parent groups. Moreover, the inclusion of the related KDO8PS and NeuB sequences in the analysis enables to identify the most likely scenarios that retraces the evolution of PEP aldolases.

2.2 Methods

2.2.1 Retrieval of sequences

In order to create hidden Markov model (HMM) profiles that accurately represent the three enzyme families, the seed sequences from the three Pfam [214] families PF00793 (*DAHP synthetase I* - which includes type Iβ and type Iα DAH7PS as well as KDO8PS), PF01474 (*DAHP synthetase II* – which includes type II DAH7PS) and PF03102 (*NeuB*) that are related

to this study were first retrieved. Each seed sequence was aligned using Clustal Ω [215] implemented in Seaview [216] and used as input file for the jackhmmer tool available on the HMMER web server [217]. Jackhmmer generated a HMM for each multiple sequence alignment (MSA) and this HMM was, in turn, used to scan four databases: UniProt [218], RefSeq [219], Pfamseq [214] and NR [219]. Initial tests showed that using an E-value sequence cut-off of 10^{-15} allowed to avoid the most obvious short sequence fragments and without losing any relevant sequences. For each database, the scan was iterated until convergence, i.e. no new sequences found. The results from each database were then grouped together and duplicates removed using Jalview [220], leading to 11591 type I DAH7PS and KDO8PS sequences, 2678 type II sequences and 3088 NeuB sequences.

2.2.2 Clustering and curation

From the sequences obtained, the identification of the different families and subfamilies is not straightforward due to the diversity of structural decorations found in these enzymes and the low overall sequence identity. In particular, type Iα, type Iβ, and KDO8PS sequences are mixed together. Therefore, to identify homogenous groups of sequences that correspond to the different families and subfamilies, the three set of sequences obtained in the previous step were combined and processed with the clustering method implemented in CLANS [221]. Specifically, after an all-against-all BLAST search of the sequences, the force directed pairwise similarities clustering algorithm was run for more than 5000 iteration cycles at a *P*-value of 10^{-40}, yielding five clusters that were identified as type Iβ, type Iα, type II, KDO8PS and NeuB and for which the sequences could be retrieved. In order to further remove sequence fragments and extremely long sequences consisting of several genes fused together, each cluster of sequences was curated based on its sequence length. More specifically the sequences that were

shorter than the mean sequence length of the cluster minus three standard deviations, or longer than the mean sequence length plus three standard deviations were discarded. This enabled outliers to be removed without tampering with the sequences corresponding to unregulated enzymes or enzymes with very long structural decorations. The average length and standard deviation of the sequences composing each cluster are reported in Table 2.1. Moreover, sequences that included the words “chimera”, “mutant”, “variant”, “synthetic”, “scaffold” within their ID were removed as they correspond to man-made sequence variants, and sequences for which the organism was “unknown” or “unnamed” were also removed.

Table 2.1. Average sequence length and standard deviation for each cluster.

	Average sequence length	Standard deviation
Type I α	358.5	20
Type I β	341	57
Type II	451	74
KDO8PS	277.5	24.5
NeuB	358	92

CLANS was then run individually on each previously identified and curated large cluster in order to obtain a finer clustering of each family and subfamily. About 5000 iteration cycles were run for each initial cluster. Identification of the different clusters was conducted automatically using the network based clustering algorithm available in CLANS with a minimum number of sequences per cluster defined at 20 and the maximum number of rounds set to 100 and refined manually when needed.

2.2.3 Selection of representative sequences and alignment

The important sequence variations complicate significantly the alignment of the different DAH7PS sequences as most alignment software aligned distinct domains such as an ACT and CM domains. Among the different algorithms tested: Clustal Ω [215] and Muscle [222] with modified gap penalty, MAFFT [223] with the E-INS-I method, T-Coffee [224] with the accurate mode and PRANK [225] with the +F option, PRANK was the only programme to correctly align the domains and extensions and was used in all the subsequent multiple sequence alignment (MSA). However, PRANK can only handle a limited number of sequences and, therefore, the alignment could not be directly performed on most of the clusters. Instead, a representative selection of sequences was extracted from each cluster following the procedure described below, leading to a number of sequences PRANK can handle without compromising the phylogenetic and taxonomic information contained in the initial set of sequences.

For each of the fine clusters identified, the sequences were exported in separate fasta files and the identifier of each sequence modified to include only the accession number and the genus and species of the organism. The taxonomy information of each sequence was then retrieved from the NCBI taxonomy database [226] and appended to the identifier of each sequences using an in-house Python programme. Taxonomic information was then used to determine to which phyla and class each DAH7PS type belongs. Sequences, which were tagged as unknown or belonging to ecological or organismal metagenomes, were discarded. Using T-Coffee [224], sequences from each fine cluster presenting more than 90% sequence identity or having less than 40% average accuracy with the other sequences were removed. For each fine cluster, the resulting sequences were aligned with PRANK [225] using the +F option over 5 iterations. The resulting alignments were then used to curate manually the sequences,

removing sequences missing the key catalytic residues or extremely long extensions only present once. Each resulting fasta file was then split into multiple fasta files; one for each phylum and a set of representative sequences was extracted from each of the resulting sequences using T-Coffee (-action +trim _seq_%%90_O40 command). The number of representative sequences selected was proportional to the total number of sequences present in each previously generated file in order to obtain a set of sequences with a good representability of each phylum. The selected sequences for each phylum for each fine cluster were then concatenated into a single file and aligned with PRANK using the +F option over 10 iterations, leading to a representative selection, both in terms of phylum and sequence differences, of aligned sequences for each DAH7PS type and KDO8PS. To remove the columns that do not contain phylogenetic information and reduce the computational cost of the calculations, Noisy [227] was used using default setting, owing to its capacity to reduce an alignment without worsening the phylogenetic results [228], and all columns with a calculated Noisy score below 1 were removed using Trimal [229]. The entire procedure, except manual curation, was automatized with python and bash scripts to ensure a robust, yet flexible, reproducible selection of representative sequences.

2.2.4 Construction of the phylogenetic trees

For each MSA, the best fitting model was determined using ProtTest3 [230] and led to LG+G (Le and Gascuel substitution model and gamma distribution to account for rate variations among sites) for type Iα, type Iβ, KDO8PS and LG+G+F (F indicate the use of empirical base frequencies drawn from the alignment) for type II. Using these models, maximum likelihood (ML) trees were constructed with RAxML 8.2.2 [231] using the rapid bootstrapping algorithm ('-f a' option) which performs tree search and rapid bootstrapping

concomitantly. 1000 rapid bootstrap searches were performed for each multiple sequence alignment (MSA) (corresponding to 200 ML tree searches). Tree construction using the conventional bootstrapping (100 replicates) and separate ML tree search (200 searches) was also tested but did not lead to significantly better trees (based on Shimodaira-Hasegawa (SH) test) and because the method was an order of magnitude slower, rapid bootstrapping was preferred. Branches of the resulting best trees presenting bootstrap support value lower than 10% were collapsed using Dendroscope [232] and the tree figures prepared with Figtree [233].

A set of representative sequences from each tree was then extracted from the phylogenetic trees and was aligned with PRANK, and the MSA reduced with Noisy as previously. The resulting MSA served as an input for a phylogenetic tree that includes all DAH7PS types as well as KDO8PS, constructed with RAxML using the same method as previously and the LG+G model suggested by ProtTest3.

In order to root the tree, sequences from NeuB were added to a selection of sequences from each MSA. However, due to the major sequences differences between NeuB, KDO8PS and DAH7PS, the alignment obtained with PRANK was not completely accurate with some conserved regions not correctly aligned. In order to overcome this, the programme BAli-Phy [234] which relies on the concomitant Bayesian inference of both the alignment and phylogenetic tree, was used. Only a limited number of sequences can be used in this method in order to ensure convergence and thus, the analysis is performed only on a small subset of the enzyme families. Eight independent MCMC chains using the LG+G+I model were run for about 55000 iterations each and the resulting samples combined using a 10% burn-in. Potential scale reduction factors, and the average standard deviation of split frequencies were calculated to be 1.003 and 0.012, indicating convergence.

2.3 Results

2.3.1 Clustering of DAH7PS, KDO8PS and NeuB sequences

To analyse the phylogenetic relationships between members of the different types of DAH7PS, and to understand how allosteric regulation has evolved within this family, the first step consisted of retrieving the sequences from a large number of proteins belonging to this family, but also to the closely related KDO8PS and NeuB families. Due to their great diversity in terms of sequences and structures, identification of sequences that belong to these families is not straightforward. Here, profile hidden Markov models (HMM) as implemented in HMMER3 [217] were used to identify remote sequence homologs that possess the PEP aldolase function found in the three families of interest. The resulting sequences were then clustered in homogenous groups of sequences using CLANS [221], which relies on the all-against-all BLAST search of an unaligned set of input sequences followed by a clustering based on a force-directed graph drawing algorithm. The result of the clustering performed at a cut-off, refer to as *P*-value, of 10^{-40} on the ensemble of DAH7PS, KDO8PS and NeuB sequences is shown in Figure 2.3. The cut-off of 10^{-40} means that the sequences having a statistical significance threshold for reporting matches against database sequences (*E*-value) higher than 10^{-40} are not subjected to the force directed algorithm of CLANS.

Five clusters can be identified: type Iβ, type Iα, type II, KDO8PS and NeuB. Moreover, type Iβ is composed of one major cluster and a smaller adjacent one, which is relatively distant. This sub-cluster corresponds to C-terminal CM regulated type Iβ. The distance between this group and the rest of the type Iβ family is intriguing, especially considering that the other subgroups (N-terminal CM domain, N-terminal ACT domain, unregulated) cannot be identified at this *P*-value level.

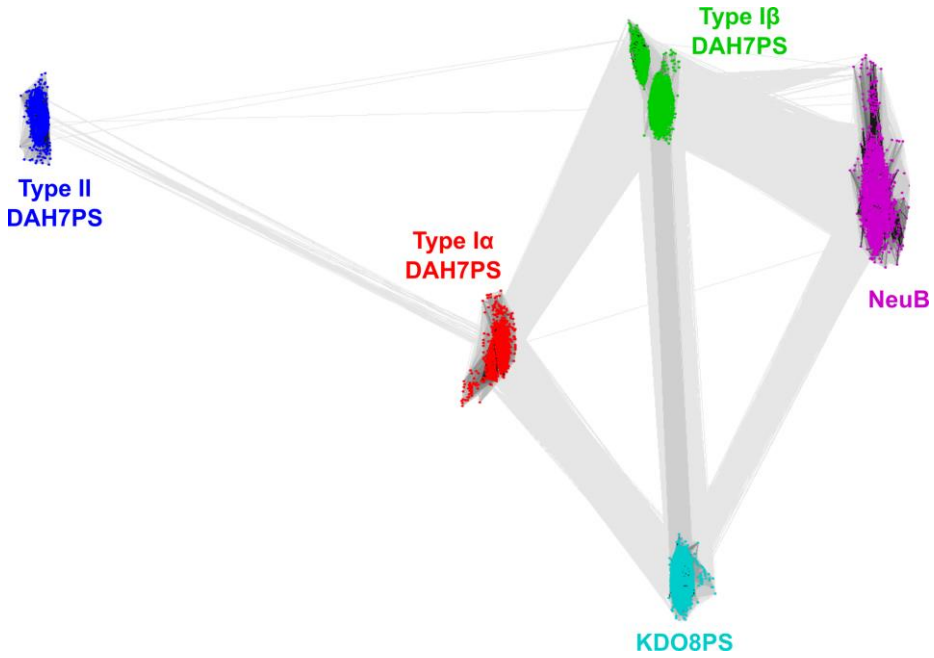


Figure 2.3. Clusters obtained after running CLAN on the ensemble set of DAH7PS, KDO8PS and NeuB sequences obtained from the HMMER searches. Type Iβ sequences are coloured in green, type Iα in red, type II in blue, KDO8PS in cyan and NeuB in purple. Grey lines represent the BLAST high scoring sequence pairs and are a measure of how similar two sequences are; here only scores having a P-value below 10^{-10} are represented with a gradient of grey (the darker the more similar the sequences are). The sequences are projected onto a Cartesian coordinate plane.

Overall this clustering provides an overview of the sequence identity between the different families and subfamilies. Type II DAH7PS sequences represent the most distant group and present very limited connections at a *P*-value of 10^{-10} with type I DAH7PS clusters, no connections with the KDO8PS and NeuB. This was also observed during the HMMER scan using the type I HMM model: with an E-value cut-off of 0.01, only 1 sequence of type II DAH7PS was picked at the first iteration and a few more after the second iteration. As a comparison, a larger number of NeuB sequences were picked up during this first two iteration steps using the same cut-off. Stronger connections exist between the type II and type Iα clusters, which suggests that type II shares a better pairwise sequence alignment with type Iα than with

type Iβ and might be more closely related. NeuB sequences are mainly connected to type Iβ and KDO8PS than with type Iα, suggesting again that NeuB is more related to type Iβ and KDO8PS. Type Iβ and KDO8PS present the strongest connections which agrees with the neighbour-joining tree presented by Jensen et al.[207].

The CLANS clustering presented here has been done on the "raw" sequences obtained from the HMMER scans and therefore, the set of sequences still contains sequence fragments and very long, multidomain sequences that are only present once and are likely the result of automated prediction based on adjacent genes rather than true multifunctional proteins. These outlier sequences were filtered based on the average sequence length of each cluster, as described in section 2.2.2. After curation, each cluster of sequences was submitted individually to CLANS in order to perform a finer clustering and identified possible subgroups and additional outliers. Clusters of sequences obtained from this finer level of clustering could then be further curated, and representative sequences extracted for phylogenetic tree construction and sequence analysis. Here, only the detailed results for the DAH7PS enzymes are presented.

The repartition of the thereby obtained DAH7PS sequences as well as their global phyla representation is shown in Figure 2.4. The majority of sequences (51%) belongs to type Iα and type Iβ and type II DAH7PS represent 25% and 24% of the sequences respectively, which is not surprising given that type Iα is often present as two or three isozymes within a single organism. In addition, 47% of the sequences belong to Proteobacteria, 18% to Firmicutes and 12% to Actinobacteria. Eukaryotes are essentially represented through Fungi (6%) and Viridiplantae (2%).

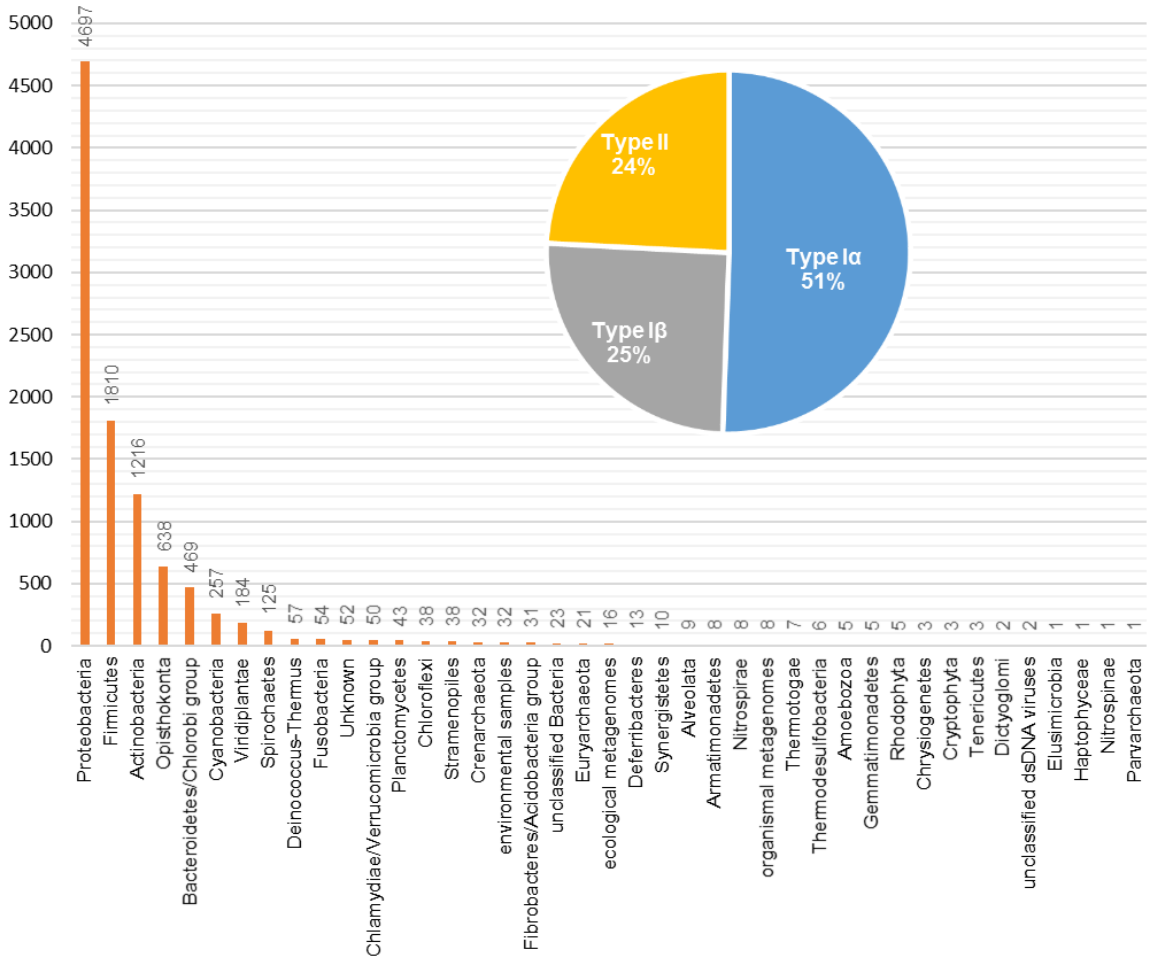


Figure 2.4. Phylum and type to which belong all the DAH7PS sequences used in the study after curation.

2.3.2 Type Iβ DAH7PS

2.3.2.1 Clustering and taxonomy

Finer clustering of type Iβ (Figure 2.5) not only confirmed C-terminal CM domain DAH7PS as being a clear subgroup but also permitted the identification of several additional clusters within the main group. Of particular interest N-terminal CM domain DAH7PS form a relatively homogeneous group in which known structures *Lmo*DAH7PS and *Gsp*DAH7PS can be found. Intriguingly, ACT-domain regulated and unregulated DAH7PS sequences form a

single group. Three minor groups were also identified with one corresponding to Leptospiraceae and another one to Chlamydiaceae.

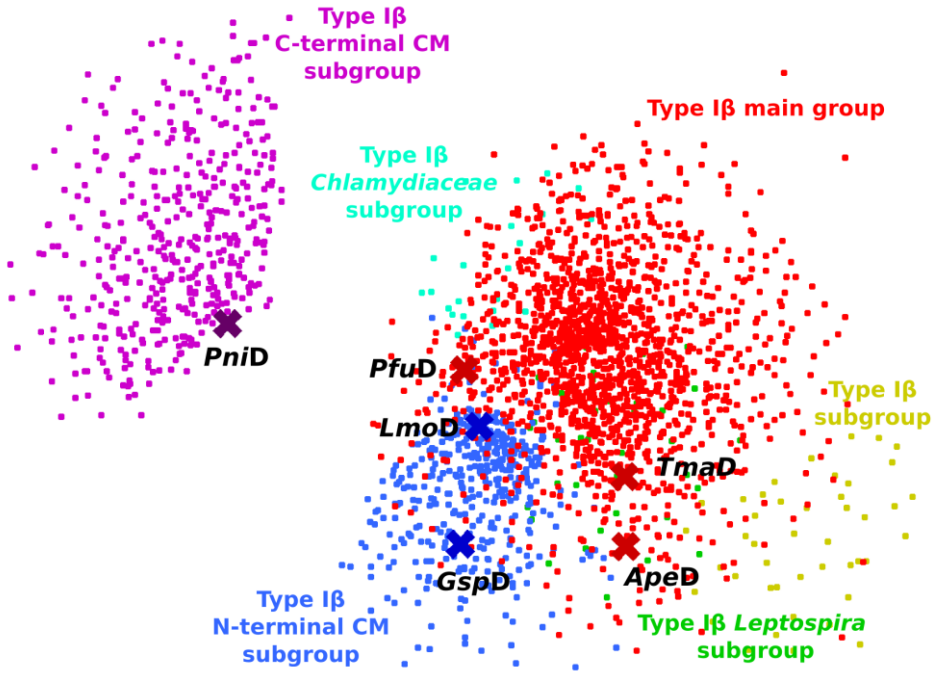


Figure 2.5. CLANS clustering of type Iβ sequences. Each dot represents a sequence and each cross corresponds to a known DAH7PS structure or an enzyme currently studied in the Parker group. *Pni*D: *Prevotella nigrescens* DAH7PS, *Pfu*D: *Pyrococcus furiosus* DAH7PS, *Tma*D: Thermotoga maritima DAH7PS, *Ape*D: *Aeropyrum pernix* DAH7PS, *Lmo*D: *Listeria monocytogenes* DAH7PS, *Gsp*D: *Geobacillus sp.* DAH7PS. The sequences are projected onto a Cartesian coordinate plane.

Analysis of organisms that possess type Iβ DAH7PS (Figure 2.6), reveals that most of the 2523 sequences belong to Firmicutes (46%) and to Bacteroidetes/Chlorobi (18%) phyla. Type Iβ DAH7PS is found in almost all bacterial phyla, with the exception of Caldiserica, Aquificae and Lentisphaerae, or, at least, was not retrieved from the databases at the time of the analysis. Moreover, type Iβ is found in Archaea, more specifically in Crenarchaeota and Euryarchaeota but not in Eukaryotes. C-terminal CM DAH7PSs are primarily found in Bacteroidetes (total of 421 sequences) and to a far lesser extent in Proteobacteria (delta-epsilon

subdivision, 9 sequences). A few other Bacteroidetes are found in the main clade (30 sequences) but do not share a C-terminal CM domain.

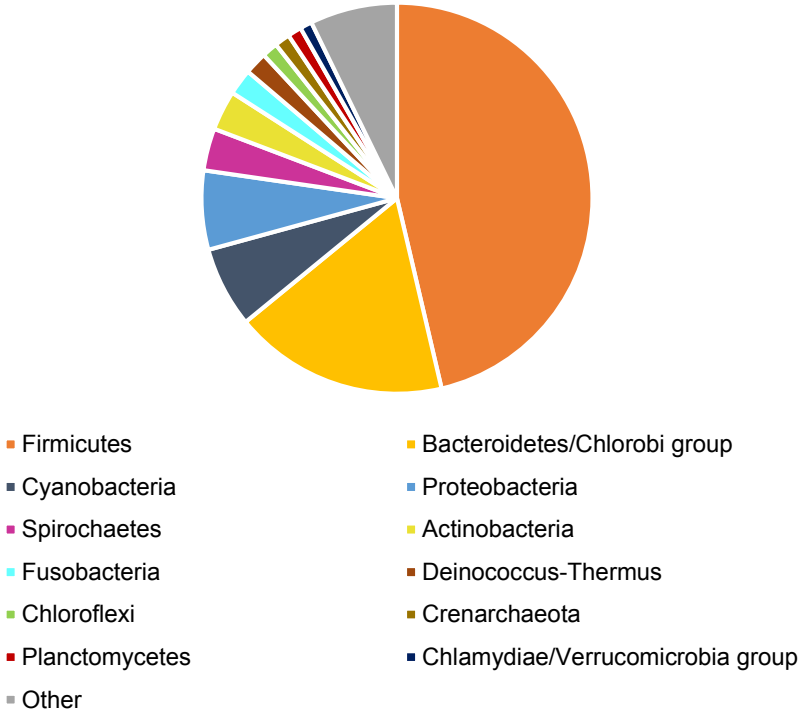


Figure 2.6. Taxonomy of the organisms in which a type Iβ sequences were identified.

2.3.2.2 Phylogeny and sequence analysis

A ML phylogenetic tree for a representative selection of 332 type Iβ DAH7PS is presented in Figure 2.7 for which branches presenting a bootstrapping value below 10% were collapsed in soft polytomies. At the deepest level of branching, two clades are identified, one corresponding to C-terminal CM DAH7PSs and the other one to the main group of type Iβ DAH7PS enzymes. This divergence, which is in good agreement with the CLAN analysis, is strongly supported by bootstrap support value. The subgroup of C-terminal CM DAH7PSs was used to root the tree as it is a valid outgroup, as shown in section 2.3.5.

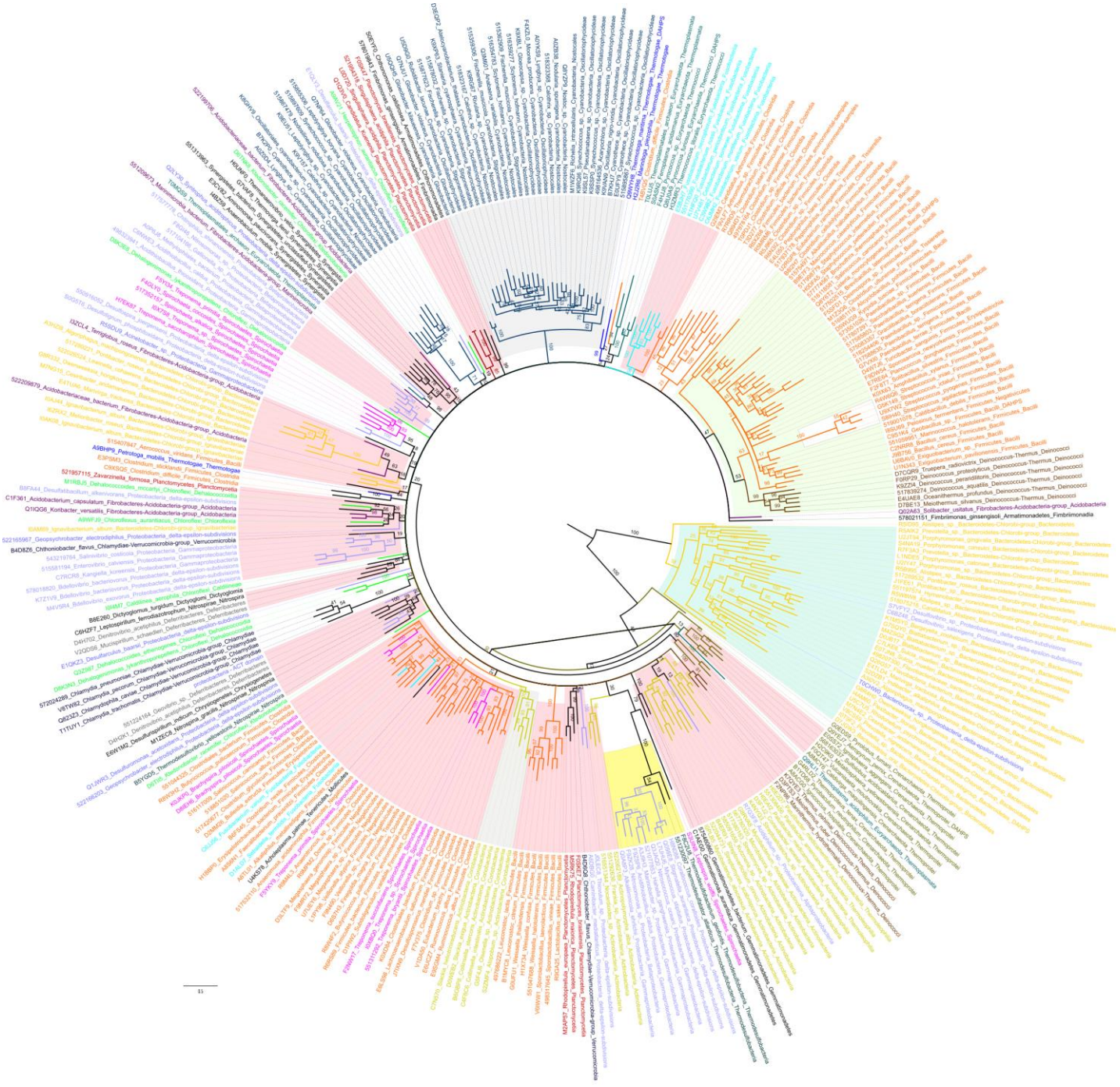


Figure 2.7. ML phylogenetic tree of type Iβ DAH7PS. The numbers on each branch correspond to the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML. Branches with bootstrap support values below 10% have been collapsed into soft polytomies. Taxa are coloured by phyla. The tree is rooted using the major subgroup of type Iβ DAH7PS (C-terminal CM DAH7PS) as an outgroup. Regions that are highlighted correspond to the taxa equipped with a regulatory domain: light pink: N-terminal ACT domain, light cyan: C-terminal CM domain, light green: N-terminal CM domain, light blue-grey: ACT domain and unregulated mixed together, light yellow: ACT domain and TIM barrel extensions.

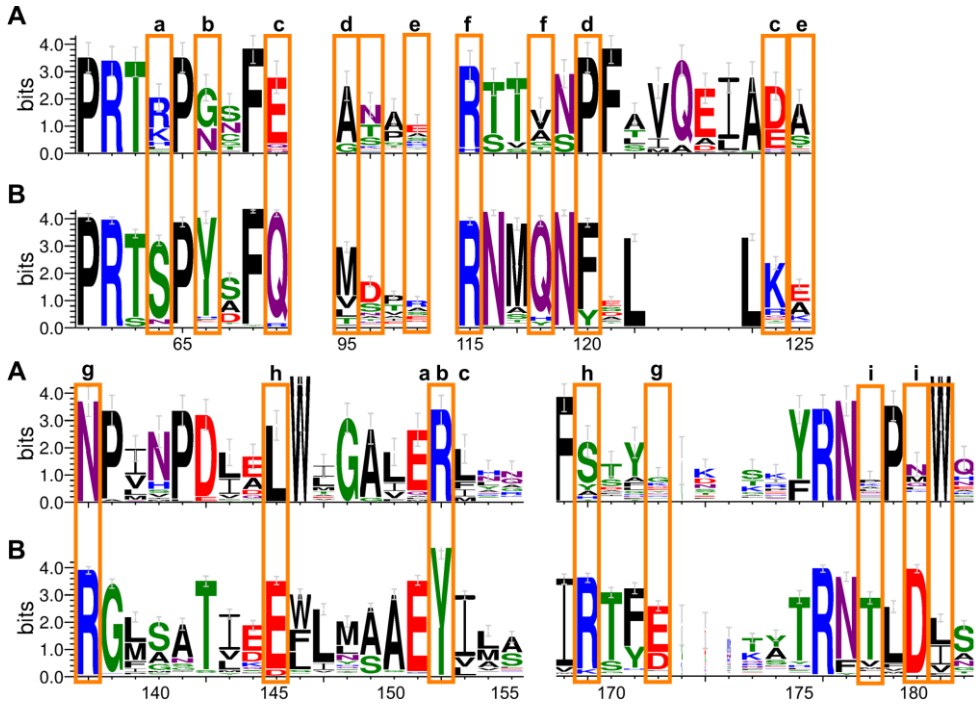


Figure 2.8. Representation of the conserved residues in the subgroup of type Iβ DAH7PS corresponding to the C-terminal CM domain DAH7PS (A) and the main group of type Iβ DAH7PS (B). The figure was obtained with WebLogo3 [235] from the alignment of representative sequences of type Iβ DAH7PS used to generate the ML tree in Figure 2.7. Residue numbering corresponds to the sequence numbering of *Pfu*D. Orange rectangles highlight the conserved residues in the sequences belonging to the main group of type Iβ DAH7PS that have been analysed as forming key non-covalent interaction between a monomeric unit and its two adjacent counterparts to form a tetrameric assembly of *Pfu*D [236]. The letter on top of some rectangles enables to identify interacting residues in the main group (e.g. Tyr152 interacts with Ser64, Tyr66 and Gln69).

The subtype of type Iβ DAH7PS corresponding to C-terminal CM DAH7PS appears to be significantly distinct from the rest of type Iβ DAH7PS (highlighted in light cyan on in Figure 2.7). In terms of sequences, it appears that most of the residues that are deemed important for the formation of a tetramer in *Pfu*D and which are conserved among sequences belonging to the main group of type Iβ DAH7PS are replaced with other residues in C-terminal CM DAH7PS (Figure 2.8) [236]. These residues tend not to favour the interactions required to form the tetrameric assembly or the canonical dimer assembly observed in the main group of type

Iβ DAH7PS. For example, using *Pfu*D residue numbering, Glu145 and Arg169 that form a strong coulombic interaction are replaced with a Leu and a non-conserved residue respectively, and the same is true for the favourable coulombic interaction between Arg137 and Glu172, which is lost. Similarly, the hydrogen bond between Gln69 and Lys124 is replaced with a coulombic repulsive interaction between a Glu and an Asp (or Glu). It has recently been confirmed experimentally with the study of *Prevotella nigrescens* DAH7PS (*Pni*D) which is equipped with C-terminal CM domain, that this loss of residues prevents this subtype of DAH7PS to form a tetramer or a classical dimer. Indeed, early results[1] based on analytical SEC and AUC suggest that *Pni*D exists as a dimer in solution (only known case for a DAH7PS to date). Moreover, based on the *ab initio* modelling of SAXS data, the dimer seems to sustain via the dimer formed by the CM domains rather than via the DAH7PS domains, and the truncated version of *Pni*D that lacks the CM domain is monomeric in solution.

Within the main group, a monophyletic group composed of Firmicutes and Deinococci sequences correspond to the N-terminal CM DAH7PS, which are not found elsewhere (highlighted in light green in Figure 2.7), as well as a group of *Streptococcus* species which, despite being the most distant taxa from the common ancestor of this group, do not have any allosteric decorations and is therefore likely composed of unregulated enzymes. N-Terminal CM regulatory domain seems, therefore, to have only evolved found in Firmicutes (mainly Bacilli) and Deinococci DAH7PS. Most residues composing the CM domain are reasonably conserved. Moreover, key residues for the CM catalytic activity [237-240] are well conserved within the N-terminal CM domain (Figure 2.9), confirming that most of these CM domains are

[1] Bai, Y; Parker, E. J., unpublished data, **2015**.

capable of converting chorismate into prephenate, although some might be more effective than others depending on which of these residues are mutated.

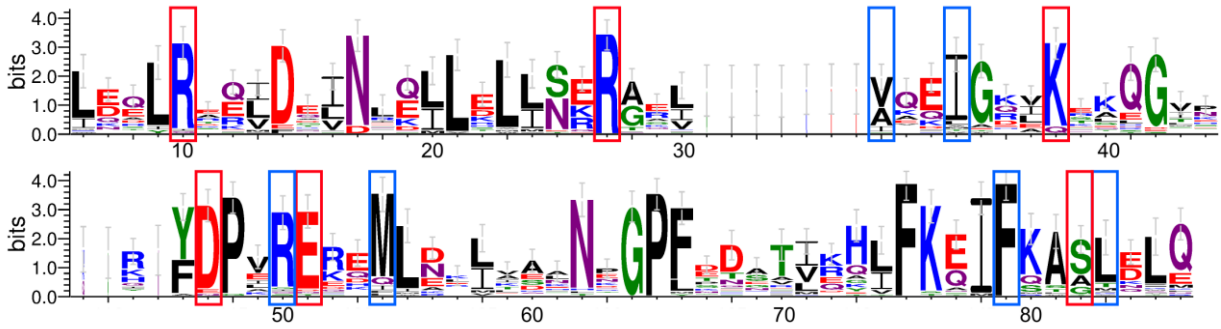


Figure 2.9. Representation of the conserved residues in the N-terminal CM domain of type Iβ DAH7PS. The figure was obtained with WebLogo3 [235] from the representative alignment of type Iβ DAH7PS sequences used to generate the ML tree in Figure 2.7. Residue numbering corresponds to the sequence numbering of *Gsp*D studied in Chapter 5. Key catalytic residues forming hydrogen bonds with the substrate chorismate or product prephenate are identified with red rectangles and those forming hydrophobic interactions with blue rectangles.

All type Iβ DAH7PSs belonging to the main group seem to share a common ancestor with two Archaea from the phylum Crenarchaeota: hyperthermophilic *Pyrolobus fumarii* and *Aeropyrum pernix* (*Ape*D), which are both unregulated. Other Crenarchaeota, which branch before the rest of the main group, present very low bootstrap support values so their phylogenetic relationship with the other DAH7PS from the main group remain unclear. Moreover, they are all equipped with a regulatory domain and cluster with an Euryarchaeota. The other archaeal Type Iβ DAH7PS, belonging to Euryarchaeota, are unregulated. Most of them, including *Pfu*D, form a clade with ACT domain regulated *Tma*D and another Thermotogae which lack a regulatory domain. This close relationship between *Pfu*D and *Tma*D might partly explain the successful ACT domain transfer from *Tma*D to the unregulated *Pfu*D [241].

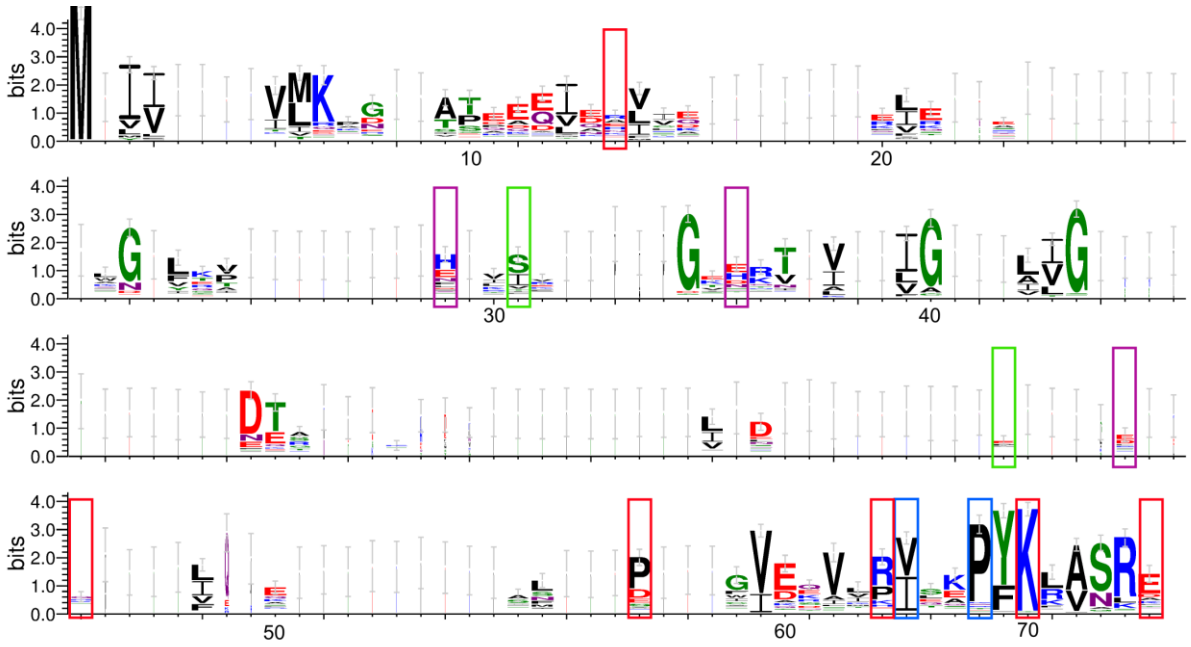


Figure 2.10. Representation of the conserved residues in the N-terminal ACT domain of type Iβ DAH7PS. The figure was obtained with WebLogo3 from the alignment of representative type Iβ DAH7PS sequences, used to generate the ML tree in Figure 2.7. Residue numbering corresponds to the sequence numbering of *Tma*D studied in Chapter 4 Region corresponding to residues 1 to 75 in *Tma*D. Key residues forming hydrogen bonds or salt bridge between adjacent regulatory ACT domain to maintain a closed form are highlighted with purple rectangles. Residues for which the side chain binds the allosteric ligand in TmaD are highlighted with green rectangles. Residues of the ACT domain interacting with residues of the catalytic main barrels are identified with red rectangles and residues playing a key role in the flexibility of the linker region are identified with blue rectangles.

In general, sequences possessing an ACT domain similar to that of *Tma*D do not form distinct clades but instead are mixed with sequences that lack a regulatory domain. Moreover, the residues composing the ACT domains are not well conserved (Figure 2.10). For instance, most of the residues deemed important for maintaining the ACT domains in a closed form (see Chapter 4) such as His29, Ser31, Glu35, Asp45, Arg46, etc. are only very weakly conserved. In contrast, residues corresponding to the linker region (after residue 63) present a better degree of conservation, suggesting that there might be more selective pressure on the residues

composing the linker region. For example, Val65 has been predicted using MD simulations to be important for the flexibility of the linker region (Chapter 4) and is well conserved or replaced with an Ile. The importance of the linker region flexibility has been recently confirmed experimentally[2], with point mutation V65P reducing the inhibitory effect of the allosteric ligand, whereas replacing Pro68, which is even more conserved, with an Ala drastically increases the flexibility of the linker.

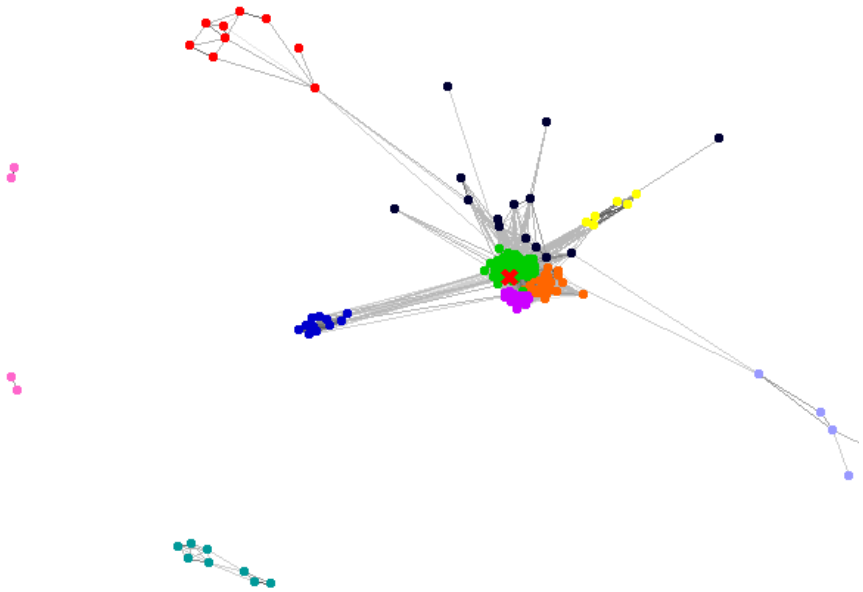


Figure 2.11. CLANS clustering of the ACT domains. ACT domain and linker region (equivalent to residue 1 to 90 in *Tma*D) from the set of representative sequences of type Iβ DAH7PS used to generate the ML tree in Figure 2.7 were used in the CLANS analysis. Grey lines represent the BLAST high scoring sequence pairs and are a measure of how similar two sequences are; here only scores having a P-value below 10^{-12} are represented with a gradient of grey (the darker the more similar the sequences are). The red cross corresponds to the ACT domain of *Tma*D. The sequences are coloured according to the cluster to which they belong. The sequences are projected onto a Cartesian coordinate plane.

[2] Moore, J.; Parker, E. J., unpublished data, **2016**.

Despite the lack of conservation between residues comprising the ACT domain, a number of patterns can be identified. Indeed, performing a CLAN analysis of the ACT domains found in the selection of sequences used to construct the tree, reveals the existence of a number of clusters (Figure 2.11). In particular, most ACT sequences are found in a group of three closely related clusters (coloured green, purple and orange in Figure 2.11) with two of these clusters corresponding to single-phylum clusters (purple and orange corresponding to cyanobacteria and Firmicutes respectively), and the biggest cluster (green) being a mix of phyla and including the ACT domain of *Tma*D. There are slightly more distant clusters such as a cluster of Bacteroidetes (yellow) and various other more isolated sequences (black). Some clusters are however much more distant and with only low connections at this level of *P*-value, accounting for the major sequences difference found in this regulatory domain, which might fold or dimerise differently in these examples to the archetypical ACT domain.

Figure 2.12. Sequence alignment of the N-terminal regulatory domains identified in a clade of Crenarchaeota. The same region in the unregulated *Ape*D is indicated at the top of the alignment for comparison.

One of these clusters (coloured in red in Figure 2.11) corresponds to two clades of Crenarchaeota equipped with an N-terminal domain containing about 60 additional residues than the N-terminal tail of *Ape*D (Figure 2.12). This N-terminal extension, which is not a CM domain does not, however, quite correspond to the canonical DAH7PS ACT domain found in

*Tma*D, as suggested by the CLAN analysis: for example, the extension of *Vulcanisaeta moutnovskia* shares only 19% identity (30% similarity) with the ACT domain of *Tma*D, whereas for *Metallosphaera yellowstonensis* it corresponds to 28% identity (37% similarity) with the ACT domain of *Tma*D. Protein domain databases Pfam [214], CATH [242] and PROSITE [243] do not return any hits for this domain. Protein secondary structure predictions for these two domains using JPred4 [244] predict a secondary structure similar to an ACT domain but with the second helix missing (Figure 2.13). On the other hand, tertiary structure predicted by homology using RaptorX [245] (Figure 2.14) results in the domain belonging to *M. yellowstonensis* being modelled as an ACT domain based on the structure of *Tma*D (*P*-value 1.03×10^{-02}), whereas in *V. moutnovskia* the regulatory domain is modelled as a αββββα motif, modelled on the basis of a DNA-binding domain (PDB:1KAF, *P*-value 9.08×10^{-03}). However, both secondary and tertiary structure predictions should be regarded with significant caution, and the actual structure of this archaeal regulatory domains might very well be a classical βαββαβ fold despite little sequence identity with the regulatory domain of *Tma*D as the ACT domain motif often present very large sequence variations and is primarily identified structurally [246].

```
M. yellowstonensis : 1---------11--------21--------31--------41--------51--------61--------71--
Sequence           : MILFVLRTGSNSSTLREKLKSTSASFRFLNLYGKEVAVAWPDQEVENVKDESVEVVVKTKKSYVLAGNEWKKDP
Jnet               : -EEEEEE----HHHHHHHHHH----EEEEEE---EEEEE-------------EEEEEE----------------
jhmm               : -EEEEEE----HHHHHHHHHH----EEEEEEE--EEEEE-------------EEEEEE-----HHHHHHH----
jpssm              : -EEEEEE----HHHHHHHHHH----EEEEEE---EEEE--------------EEEEEEE----EE---------

V. moutnovskia     : 1---------11--------21--------31--------41--------51--------61--------71--
Sequence           : MIIRASNNRVNDVASLLDRAGIKLNIVRIYGEELIVTWPDSRVDVNSVKSVDPNAVIIEVKTKYQLASNTWRDR
Jnet               : -EEEEE---HHHHHHHHH----EEEEEEE--EEEEE-----------------EEEEE----E---HHHH--
jhmm               : -EEEEE---HHHHHHHHH-----EEEEEEE--EEEEE-----------------EEEE---------HHHH--
jpssm              : -EEEE----HHHHHHHHHH---EEEEEEE----EEE------------EE-----EEEEEE--EEE--------
```

Figure 2.13. Secondary structure prediction for two N-terminal domain extensions found in Archaea *M. yellowstonensis* and *V. moutnovskia*. Predictions made with JPred4 [244]. Jnet: consensus prediction, jhmm: HMM profile based prediction, jpssm: position specific scoring matrix based prediction, H: helix region, E: β-strand region.

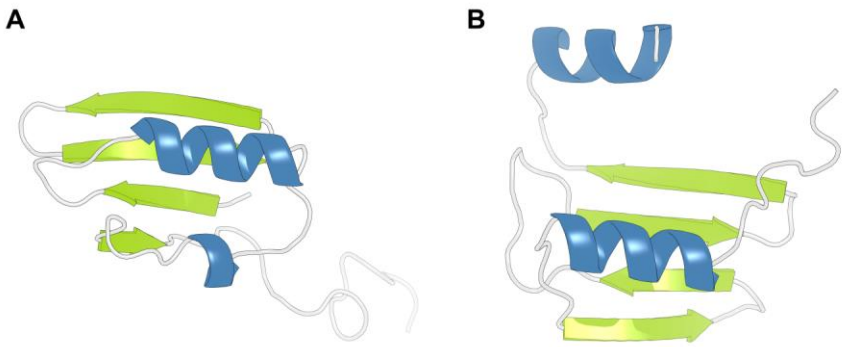


Figure 2.14. Tertiary structure prediction for two N-terminal domain extensions found in Archaea *M. yellowstonensis* (A) and *V. moutnovskia* (B). Models were generated with RaptorX [245].

One of the most intriguing group of sequences within the main group correspond to the clade formed by a number of delta-epsilon Proteobacteria and Thermodesulfobacteria and that have undergone the largest number of changes from the common ancestor of all type Iβ DAH7PS (clade highlighted in yellow in Figure 2.7). Sequences from this clade possess a longer C-terminal tail, but more surprisingly appear to have an extended $\alpha_3\beta_4$ loop and have what looks like a longer $\beta_6\alpha_6$ loop. Almost no other sequences in this group present insertions within the catalytic domain, which are a feature of type Iα and type II DAH7PS. In addition, some of the most conserved residues of the catalytic barrel, which are not involved in the catalytic reaction, are mutated. Moreover, the N-terminal ACT domain of these taxa is unconventional as it is longer than the ACT domain of *Tma*D (blue cluster in Figure 2.11) and present charged or bulky polar residues in lieu of small hydrophobic residues at some key points of the ACT domain, such as conserved Glu, Tyr and Arg residues on the β-sheet of the ACT domain, instead of a Val38, Gly40 and Ser61 respectively in *Tma*D. Modelling of the structure of one of these taxa, the DAH7PS from *Nitrosococcus oceani*, by homology with *Tma*D using Modeller for the tetrameric structure, permits the structural location of the differences to be highlighted (Figure 2.15). The insertions within the catalytic barrel are mostly

localised in the solvent exposed region of the tetramer and are not expected to affect the tetrameric assembly. Although the role of these extensions remains unclear, the mutations in the ACT domain take place at the interface between the two facing β-sheets that form the allosteric binding site. The presence of the charged or polar bulky residues is likely to affect the stability of the ACT domain interface as well as the selectivity for the allosteric ligand.

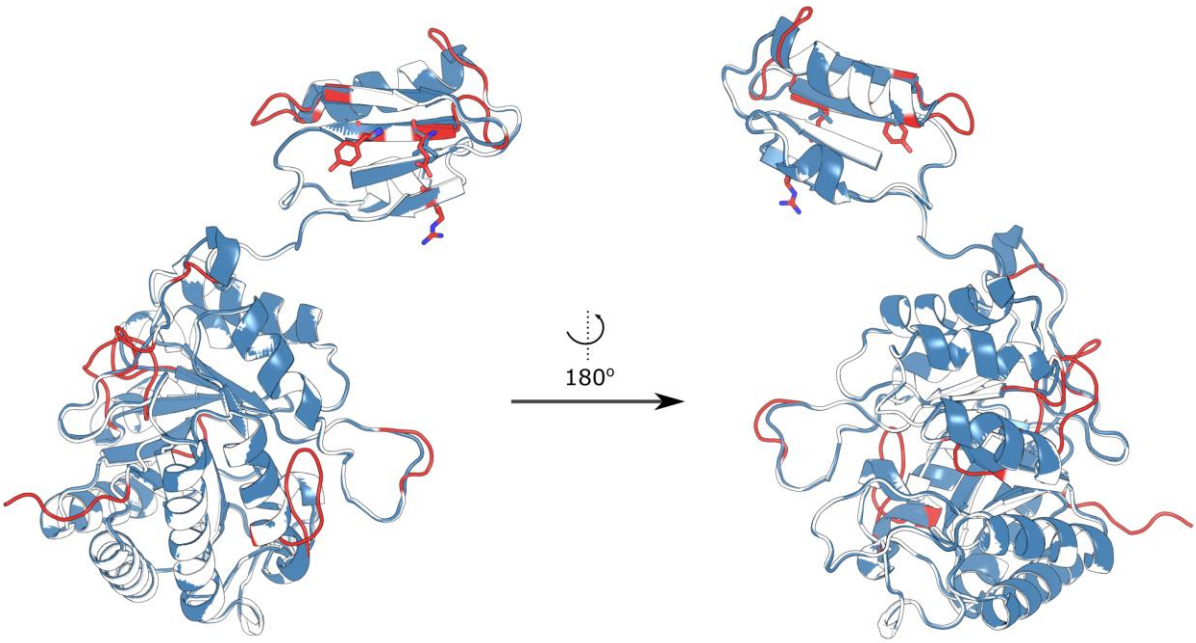


Figure 2.15. Structural model of *Nitrosococcus oceani* obtained by homology with *Tma*D. The structure of *Tma*D is represented in blue whereas the superimposed structure of the model for *N. oceani* is in white with the regions that are predicted to change in red. Major differences between residues situated on the β-sheet of the ACT domain are shown with sticks.

Surprisingly, the clade of Bacteroidetes found in the main group, belongs to the same organisms and strains than some of the Bacteroidetes found in the C-terminal CM domain subgroup. To the best of our knowledge, this represents the first noted example of isozymes within type Iβ DAH7PS. Members of this clade are equipped with an N-terminal domain that resembles the sequences of the N-terminal ACT domain described in the above paragraph although shorter. This group corresponds to the yellow cluster Figure 2.11. Based on what is

known for the CM-regulated and ACT-regulated enzymes, the ACT domain isozyme might be sensitive to the aromatic amino acids whereas the CM-regulated isozyme might be inhibited by chorismate and prephenate.

Overall the topology of the tree suggests that C-terminal regulated type Iβ DAH7PS on one hand and N-terminal regulated along with unregulated type Iβ DAH7PS on the other hand, evolved from a common ancestor. Evolution of the N-terminal CM domain from the last common ancestor of the unregulated and N-terminal type Iβ DAH7PS is also well supported and it seems that a lineage of Streptococcus may have lost the N-terminal CM domain through evolution. However, taxa equipped with an N-terminal ACT domain or that do not have regulatory extensions are mixed together, without clear evolutionary or taxonomic pattern. This observation might suggest that the evolutionary pressure is more important on the catalytic barrel than on the regulatory ACT domains. This is confirmed by constructing an equivalent tree as in Figure 2.7, but following removal of all the allosteric terminal domains. The topology of this tree (Figure 2.16) is very similar to that of the tree with the regulatory domain, in particular for the unregulated and ACT-domain regulated lineages.

Figure 2.16. ML phylogenetic tree for the catalytic barrel of type Iβ DAH7PS. The numbers represent the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML. Branches with bootstrap support values below 10% have been collapsed into soft polytomies. Taxa are coloured by phyla. The tree is rooted using the major subgroup of type Iβ DAH7PS (C-terminal CM DAH7PS) as an outgroup.

2.3.3 Type Iα DAH7PS

Clustering of type Iα sequences using CLAN (Figure 2.17) enables the identification of a major subgroup (type Iα subgroup) distant from the rest of the main group. The main group itself can be subdivided into 5 sub-clusters, including one of *Pseudomonas* species and another one of *Actinobacteria* species. All the known type Iα DAH7PS structures belong to the same cluster of the main group.

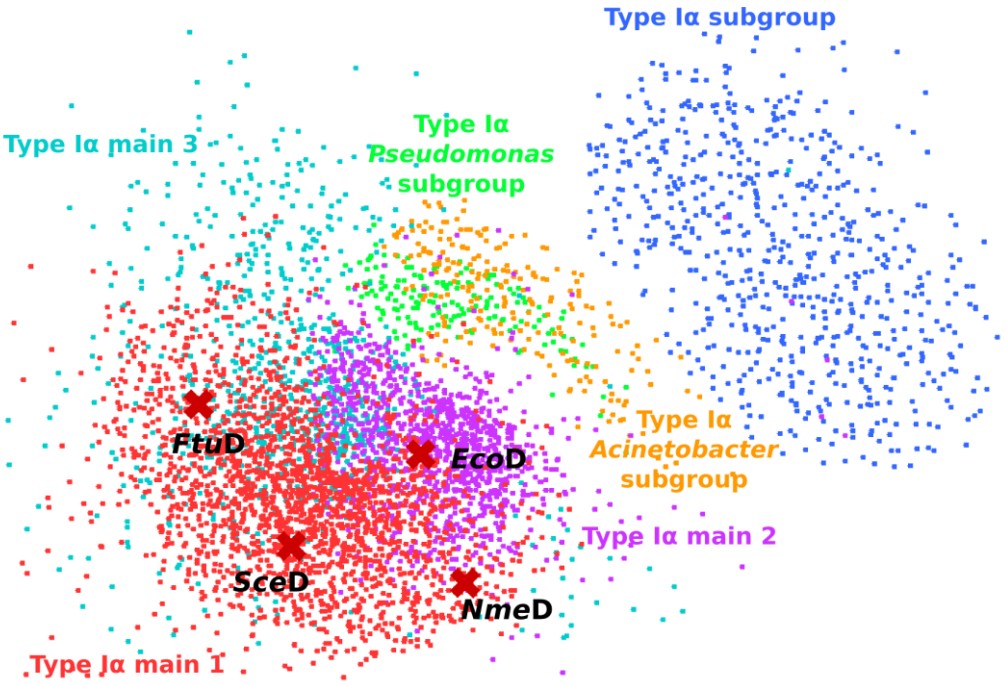


Figure 2.17. CLAN clustering of type Iα DAH7PS sequences. Each dot represents a sequence and each cross corresponds to a known DAH7PS structure or an enzyme currently studied in the Parker group. *Ftu*D: *Francisella tularensis* DAH7PS, *Nme*D: *Neisseria meningitidis* DAH7PS, *Sce*D: *Saccharomyces cerevisiae* DAH7PS, *Eco*D: *Escherichia coli* DAH7PS. The sequences are projected onto a Cartesian coordinate plane.

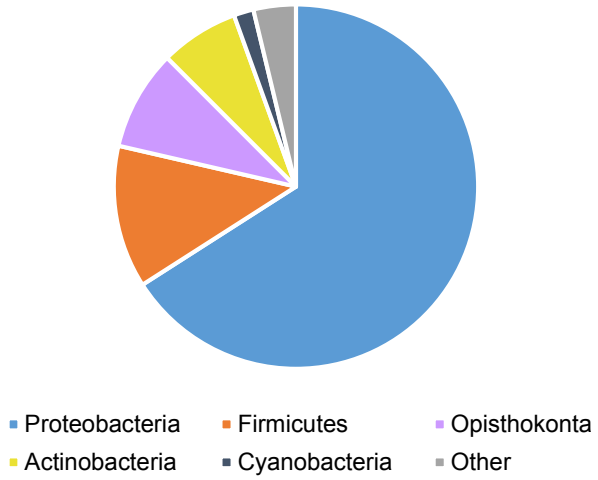


Figure 2.18. Taxonomy of the organisms in which a type Iα sequences were identified.

Analysis of the organisms in which type Iα DAH7PS is present reveals a narrower taxonomic diversity when compared with type Iβ DAH7PS. Type Iα DAH7PS is mainly found in four bacterial phyla: Proteobacteria, which is by far the most common in this DAH7PS type (66% of the 5048 type Iα DAH7PS sequences identified for this analysis), Firmicutes (13%), Actinobacteria (7%) and some Cyanobacteria (2%). The other bacterial phyla, although negligible in proportion can nonetheless represent a substantial number of sequences reaching sometimes the same level than in type Iβ: for example, the same number of sequences belonging to the Chlamydiae/Verrucomicrobia group in both subtype. In contrast to the type Iβ DAH7PS, in which was found in a number of Archaeal sequences, none are represented here, but instead, type Iα is present in Eukaryotes, especially in Fungi (9%), Stramenopiles (0.4%), Amoebozoa, Alveola, Cryptophyta (0.1% each).

Figure 2.19. ML phylogenetic tree of type Iα DAH7PS sequences. The number on each branch represents the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML. Branches with bootstrap support values below 10 have been collapsed into soft polytomies. Taxa are coloured by phyla. The tree is rooted using the major subgroup of type Iα DAH7PS as an outgroup.

An ML tree for type Iα DAH7PS is presented in Figure 2.19. As for type Iβ DAH7PS, a well-supported clade, corresponding to the subgroup identified during the clustering, forms a distinct outgroup from the main group type Iα DAH7PS and was used to root the tree. Except for a few taxa, this subtype of type Iα is almost exclusively found in Firmicutes and this phylum is not present in the main group. Moreover, Bacilli in this subgroup form a distinct clade and include a number of Streptococcus species, including *Streptococcus agalactiae*, which is also found in type Iβ DAH7PS within the group of Bacilli having lost their N-terminal CM domain (although it corresponds to a different strain). Even more interestingly, sequence analysis of this subtype reveals clear differences in the regions corresponding to the N-terminal tail, β_0 and β-hairpin of type Iα DAH7PS and residues that have been identified as being responsible for the allosteric regulation of *Nme*D (Chapter 3) and which are well conserved within the main group are not present in the subgroup. In particular, using the residue numbering of *Nme*D, residues Glu98, Asp148 and Glu176 which have been shown to communicate the allosteric signal [247], as well as Asp8, Asp9, Gln153 and Ser182 that bind the amino acid function of the allosteric ligand, are all very conserved within the main group, but not conserved at all within the subgroup (Figure 2.20). The region corresponding to the second β-strand forming the β-hairpin in the main group (residues 218 to 222) is missing in the subgroup but residues composing β_6 are nonetheless conserved. Instead, there is a 5-residue insertion in what corresponds to the $\beta_6\alpha_6$ loop of the main group. Finally, residues Glu27 and Arg126, which have been shown to be essential to maintain the tetrameric structure [248], are not particularly well conserved within the subgroup. Major differences with the main group in terms of allosteric regulation and perhaps in the quaternary structure are therefore to be expected for this subgroup.

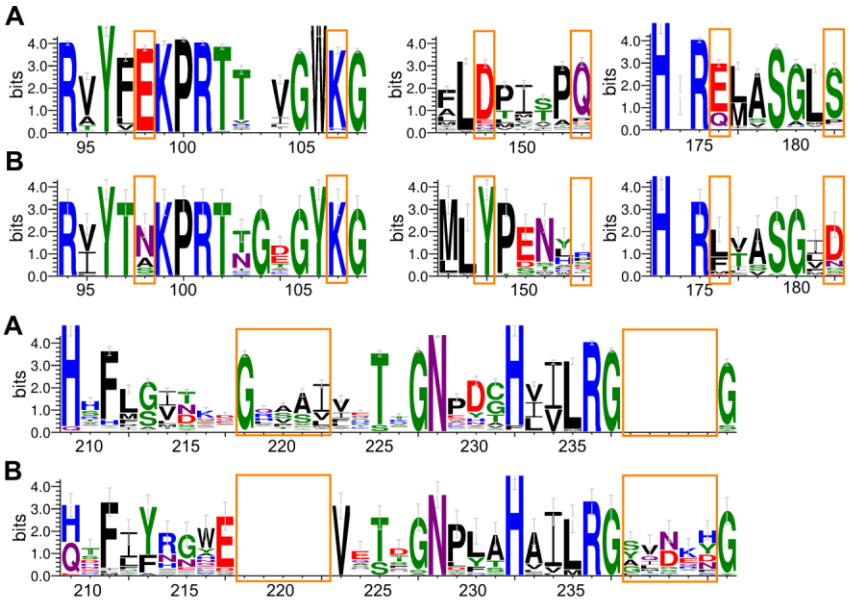


Figure 2.20. Representation of the key conserved residues in the subgroup of type Iα DAH7PS. Conserved residues in the main barrel of the main group of type Iα DAH7PS (A) and the subgroup of type Iα DAH7PS (B) obtained with WebLogo3 [235] from the alignment representative of type Iα DAH7PS used to generate the ML tree in Figure 2.19. Sequence numbering corresponds to *Nme*D (studied in Chapter 3). Orange rectangles highlight the residues of importance discussed in the main text.

Within the main groups, most residues are well conserved and no major insertions compared with the structurally known type Iα DAH7PS are observed. The main differences between sequences of the main group appear to be at the N-terminal tail, for which composition and length vary greatly from one taxon to another. Some sequences present, for example, an N-terminal tail 69 residues longer than the N-terminal tail of *Nme*D. In case of such a length, the N-terminal tail is likely to adopt a secondary structure. Querying protein domain databases Pfam [214], CATH [242] and PROSITE [243] with the sequences from the two longest N-terminal tails (found in *Metascardovia criceti* and *Bifidobacterium gallicum*) did not return any known domains. Secondary structure predictions using JPred4 [244] are shown in Figure 2.21 and propose a two helical regions and one β-strand region fold for the two tails (although not

in the same order for both these sequences), however, this is inference must be treated with caution as secondary structure prediction can be unreliable.

```
M. criceti : 1---------11--------21--------31--------41--------51--------61--------71
Sequence   : MSDIDSWSSSISDNSSAEASSVEFNSTEALERHAQEAARHRAALHAQLAQSSEQNPLDSVSRWEDEVGLARI
Jnet       : ------EEEE-----------HHHHHHHHHHHHHHH---------HH-------------------------
jhmm       : -----EEEEEE----------HHHHHHHHHHHHHHHHHH-------HHHHHH-----EEE-------------
jpssm      : --------------------------HHHHHH--------HHH-----------------------------

B. gallicum : 1---------11--------21--------31--------41--------51--------61--------71----
Sequence    : MNRQQDELSSQPDASRMNVAATAATAADHIHHVDDPSRVDPQKLHAVHQVLQEGGDPFTELTDLPRWEDQVGIDRI
Jnet        : --------------------HHHHHHHHHHHH-------HHHHHHHHHHHHHH-----EEE--------------
jhmm        : ----------------HHHHHHHHHHHHHHHHHH------HHHHHHHHHHHHHH-----EEEE-------------
jpssm       : ---HHHH-----------------HHHHH----------HHHHHHHHHHHHH-----------------EE----
```

Figure 2.21. Secondary structure prediction for the two longest N-terminal tails. The longest tails were identified in the list of representative sequences used to generate the ML tree in Figure 2.19, obtains with JPred4 [244]. Jnet: consensus prediction, jhmm: HMM profile based prediction, jpssm: position specific scoring matrix based prediction, H: helix region, E: β-strand region.

A few type Iα DAH7PS, that were indicated as belonging to animals were retrieved during the HMM search of the database. Additional database searches revealed that most of the other enzymes of the shikimate pathway had been reported for these organisms. In order to ascertain if these sequences were due to sample contaminations during the sequencing or were indeed animal DAH7PS, they were included in the phylogenetic tree construction. All the putative animal sequences are found in bacterial clades, whereas all the other Eukaryotes form a distinct clade, strongly suggesting a bacterial contamination of the sequenced animal genomes. This is not unprecedented in the study of DAH7PS [209] and a review of the literature confirmed the contamination. Indeed, *Pantholops hodgsonii* (Tibetan antelope) is found within an alphaproteobacteria clade in the type Iα tree, and as a sister taxon of an alphaproteobacteria clade in the type II tree (section 2.3.4), in agreement with Laurence and co-workers who recently demonstrated a contamination of this particular genome by the alphaproteobacteria of genus *Bradyrhizobium*. In the case of *Bombus impatiens* (common eastern bumblebee), the

DAH7PS sequence most certainly belongs to a gammaproteobacteria from the gut microbiota of this insect [249], explaining why this sequence groups with other gammaproteobacteria. Finally, the genome of *Ceratitis capitata* (Mediterranean fruit fly) has also been shown to be contaminated, with for example *Enterobacteriaceae* from its gut [250], which is consistent with the sister taxa of this DAH7PS sequence corresponding to *salmonella enterica* sequences.

True eukaryotes sequences are grouped within a single clade presenting, however, a low bootstrapping value. The subsequent branching to form clades of single phyla are, however, well supported and lead to distinct monophyletic groups of Fungi, Amoebozoa, Stramenopiles and Cryptophyta. Due to the low support value, it is not clear from which bacteria Eukaryotes DAH7PS are related to, however, the tree topology seems to suggest Proteobacteria.

2.3.4 Type II DAH7PS

CLAN clustering of type II DAH7PS (Figure 2.22) enables a number of relatively well-defined clusters to be identified. As for the two type I DAH7PS subtypes, type II also seems to have a main subgroup, which is more distant from the main group of clusters, but this subgroup is scattered and does not contain many sequences. This subgroup includes the short variant of type II DAH7PS of *Pseudomonas aeruginosa* (*Pae*D). *Mtu*D is found in a distinct cluster to the long variant of *Pae*D, whereas *Helicobacter pylori* DAH7PS (*Hpy*D) belongs to a cluster of *Helicobacter pylori* strains.

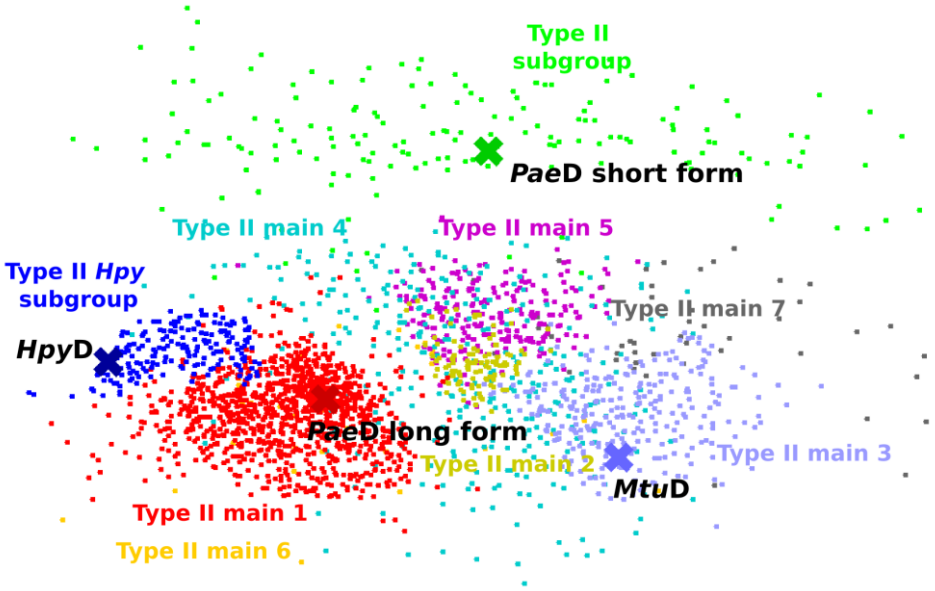


Figure 2.22. CLAN clustering of type II DAH7PS sequences. Each dot represents a sequence and each cross corresponds to a known DAH7PS structure or an enzyme currently studied in the Parker group. *Pae*D short form: short variant of Pseudomonas *aeruginosa* type II DAH7PS, *Pae*D long form: long variant of Pseudomonas *aeruginosa* type II DAH7PS. *Mtu*D: M*ycobacterium tuberculosis* DAH7PS, *Hpy*D: *Helicobacter pylori* DAH7PS. The sequences are projected onto a Cartesian coordinate plane.

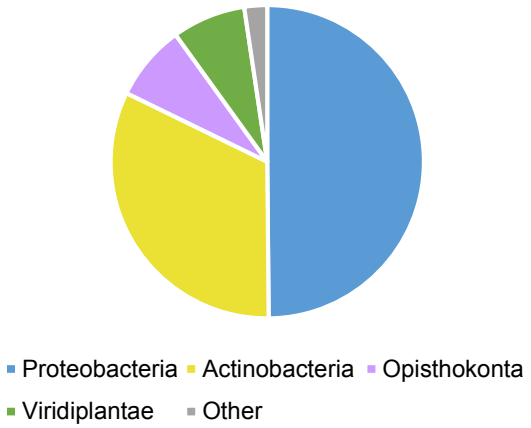


Figure 2.23. Taxonomy of the organisms in which a type II sequences were identified

Type II DAH7PS is found in even less bacterial phyla than type Iα DAH7PS (Figure 2.23), with 50% of the 2413 sequences retrieved belonging to Proteobacteria and 32% to Actinobacteria. A few Firmicutes, Bacteroidetes, Acidobacteria and some others are also present but this remains anecdotal (only 13 sequences in total). A little more than 16% of the

sequences belong to Eukaryotes, essentially fungi and plants. Type II is the only DAH7PS enzyme found in plants. To a far lesser extent, it has also been identified in Stramenopiles (16 sequences), Alveolata (6 sequences), Rhodophyta (5 sequences).

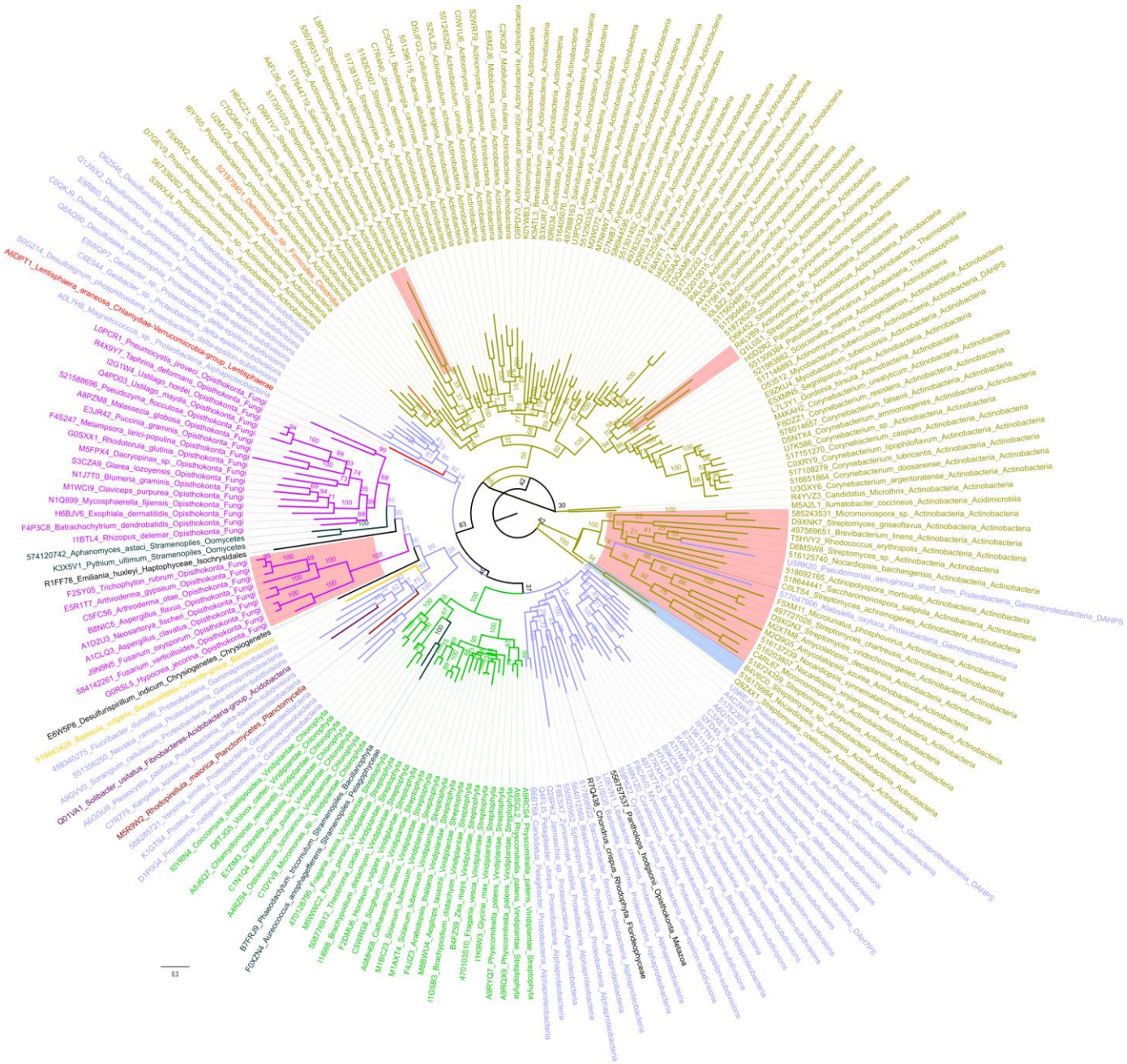


Figure 2.24. ML phylogenetic tree of type II DAH7PS. The number represent the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML. Branches with bootstrap support values below 10% have been collapsed into soft polytomies. Taxa are coloured by phyla. The tree is midpoint rooted. Regions

highlighted in light red correspond to sequences missing the $\alpha_{2a}\alpha_{2b}$ insertion, whereas the taxa highlighted in blue have an insertion different from the two $\alpha_{2a}\alpha_{2b}$ helices in the same region.

An ML tree of type II DAH7PS obtained for a representative set of sequences is presented in Figure 2.24. The subgroup identified with CLAN, to which *Pae*D short form belongs, corresponds to a well-supported clade of sequences missing the $\alpha_{2a}\alpha_{2b}$ helices responsible for the binding of Trp. This missing insertion is however not limited to this particular clade, and this couple of helices is also absent in a small number of Actinobacteria belonging to the main group, as well as in a complete clade of Fungi (highlighted in light red Figure 2.24). Moreover, within this subgroup, two taxa (highlighted in light blue Figure 2.24) present an extension in this region, which is however significantly different from the $\alpha_{2a}\alpha_{2b}$ extension sequences.

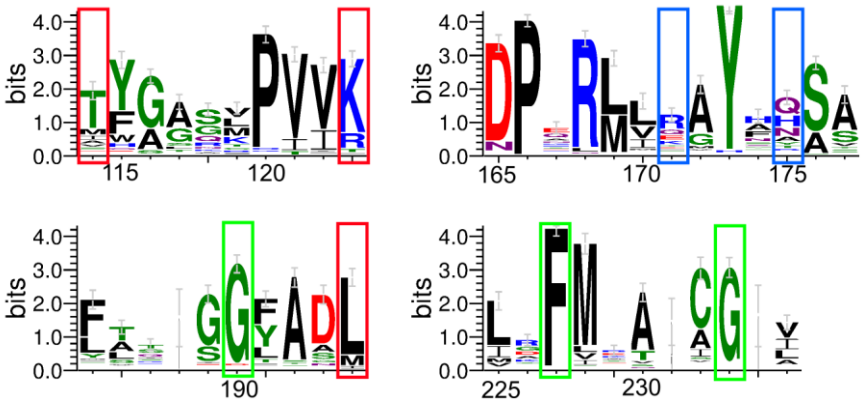


Figure 2.25. Representation of the key conserved residues in the main group of type II DAH7PS. Figure obtained with WebLogo3 from the alignment representative of type II DAH7PS used to generate the ML tree in Figure 2.19. Sequence numbering corresponds to *Mtu*D. Red rectangles highlight the residues of importance for the binding of Trp, whereas residues important for the binding of Phe are highlighted with blue rectangles and those important for the communication of the allosteric signal are highlighted in green.

Within the main group, residues that have been shown to be fundamental for the transmission of the allosteric signal or for the tetrameric assembly such as or Gly190, Phe227

and Gly232 or in the binding of Trp such as Lys123, Leu194 are particularly well conserved (Figure 2.25). In contrast, some residues deemed important for the binding of Phe such as Arg171 [187] or Asn 175 are significantly less conserved, which might indicate substantial changes in binding affinity at those sites.

Eukaryote DAH7PS are, as for type Iα, closely related to Proteobacteria. Eukaryotes, however, do not form a distinct clade but form two clades. Plant DAH7PS, as well as a few Stramenopiles DAH7PS, form a first clade, which is strongly supported by bootstrapping. Within this clade, Chlorophyta and Streptophyla are first found to diverge into two separate lineages, and then Stramenopiles diverge from Chlorophyta. The second clade consists in Fungi, Isochrysdales and Oomycetes. From the last common ancestor of this clade a first split led to a lineage of Fungi lacking the $\alpha_{2a}\alpha_{2b}$ extension, whereas, within the rest of this clade, Isochrysdales and Oomycetes diverged from the rest of the Fungi. Both eukaryotic clades have proteobacterial sister groups resulting in two larger clades. These clades share a common ancestor with a clade of Proteobacteria, which in turn share a common ancestor with Actinobacteria. Therefore, Actinobacteria share a common ancestor with any other organism equipped with type II DAH7PS, on the basis that the mid-point rooting of this tree represents a good approximation of the actual root.

2.3.5 Toward a complete DAH7PS phylogenetic tree

A set of representative sequences were selected from each of the trees detailed above in order to keep the phylogenetic diversity as well as the main tree topology. The same procedure was performed for KDO8PS. The resulting 158 sequences were then collated and aligned using PRANK as previously and the columns containing no useful phylogenetic information

removed and an ML tree constructed using the same procedure as previously. The resulting unrooted tree is presented in Figure 2.26.A.

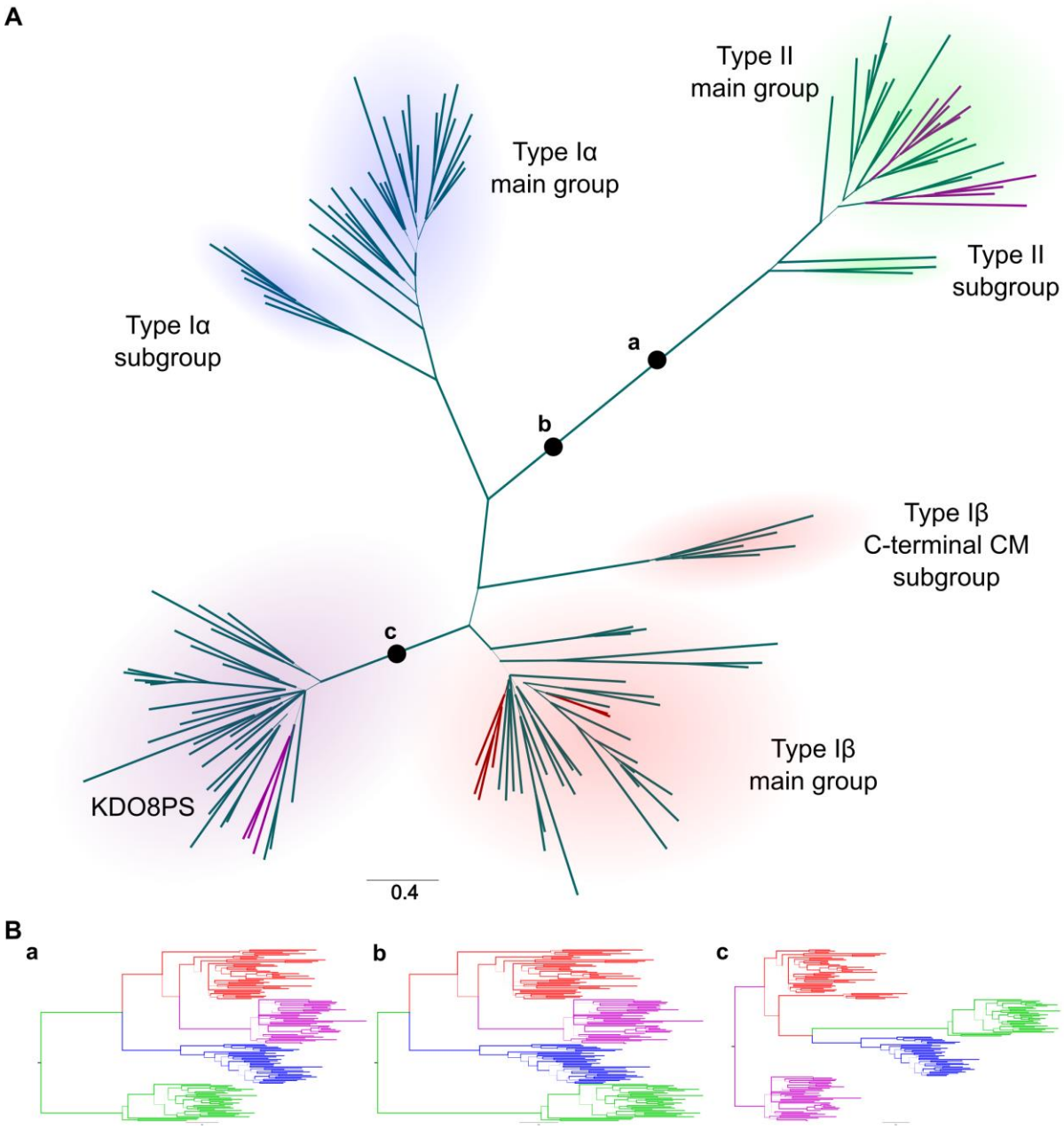


Figure 2.26. A. ML tree for a selection of DAH7PS and KDO8PS.The width of the branches is proportional to the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML. Branches with bootstrap support values below 10 have been collapsed into soft polytomies. Taxa are coloured by kingdom: dark cyan for Bacteria, dark red for Archaea and dark purple for Eukaryotes. The main clades of the tree are highlighted depending on which group they belong: red for type Iβ, blue for type Iα, green for type II and purple for KDO8PS. Circles identified with letters indicates the possible position for rooting the tree. B. Different rooted versions of the tree. The resulting trees for the different rooting points using the same type-associated colour code. (a)

corresponds to a midpoint rooted tree, (b) to a tree rooted using type II as an outgroup, (c) to a tree rooted using KDO8PS as an outgroup.

Firstly, this tree shows that in the case of type Iα and Iβ, the subgroups identified previously are robust outgroups when compared to the main group, confirming *a posteriori*, the validity to using them to root the type Iα and Iβ tree. The situation is less clear for type II, however, the subgroup is the first to diverge, validating the conclusion drawn with the mid-point rooted tree.

Rooting of this tree is not straightforward and a number of possibilities can be envisaged. Three of the possible scenarios are shown in Figure 2.26.B. Classically [207], KDO8PS, type Iα and type Iβ are studied together while type II is considered separately based on low sequence identity between type II and the other types. Following this logic, type II can be considered as an outgroup and used to root the tree, resulting in the rooted tree (a). However, such a tree results in a situation where type II is the closer to the root and has undergone less evolutionary changes than the other groups. This seems unlikely as the allosteric machinery is much more complex in type II than in type Iβ which can even be unregulated. A better rooting would, therefore, be a mid-point rooting that would better account for the evolutionary changes of type II, however, this tree presents two limitations. Firstly, almost the amount of evolutionary change from the root is observed for type Iβ and type Iα, this again would be surprising for the same reason as for type II. Secondly, it relies on the assumption that the rate of evolution is reasonably constant, which is unknown. Another option might be to consider KDO8PS as the outgroup, as it does not have the same function and is deprived of any allosteric regulation mechanism. It might, therefore, be reasonable to assume that KDO8PS diverged from DAH7PS before the apparition of allosteric regulation in the later.

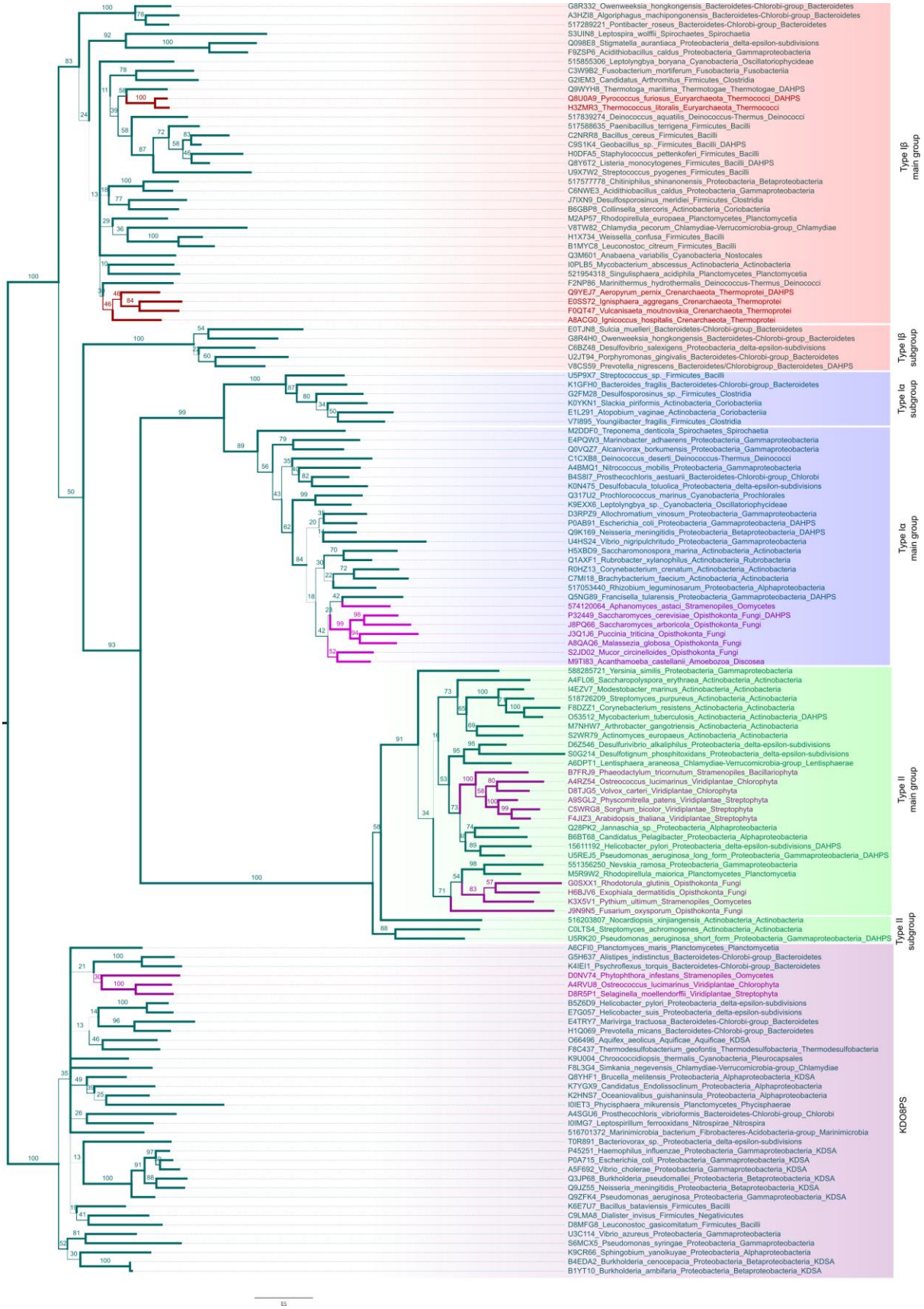


Figure 2.27. ML tree for a selection of DAH7PS rooted using KDO8PS as an outgroup. The width of the branches is proportional to the bootstrap support values obtained for 1000 rapid bootstrap replicates with RAxML.

Branches with bootstrap support values below 10 have been collapsed into soft polytomies. Taxa are coloured by kingdom: dark cyan for Bacteria, dark red for Archaea and dark purple for Eukaryotes.

A more detailed version of the tree rooted using KDO8PS as an outgroup is given in Figure 2.27. This tree presents the advantage of accounting for the amount of evolutionary changes expected from the known allosteric regulation in the different types of DAH7PS, with type II having undergone more changes than type Iα, which in turn has undergone fewer changes than type Iβ. In this tree, type Iα and type II have diverged from a common ancestor. This common ancestor has evolved from an ancestor shared with C-terminal regulated type Iβ, which is a lineage of type Iβ. Some inconsistencies between this tree and the trees obtained for each type can be noted. In particular, Fungi in type Iα and type II cluster with bacterial taxa instead of being distinct monophyletic groups. In type II, Proteobacteria and Eukaryotes do not share a common ancestor with Actinobacteria belonging to the main group of type II. Such kind of discrepancy might be due to the limited number of sequences or from the alignment. Indeed, major differences exist between DAH7PS and KDO8Ps and between the different DAH7PS types and the alignment can be challenging.

In order to improve the alignment and also to try to include NeuB sequences as an outgroup within the DAH7PS and KDO8PS, another approach was tried that relies on using Bayesian inference to concomitantly perform the alignment and the tree construction. This approach, made possible with BAli-Phy [234], has been shown to provide lower alignment errors [251] and has been successfully used in the case of highly divergent sequences [252]. However, to ensure convergence, BAli-Phy can only be used on a very limited number of sequences. This method was used on 31 taxa and the resulting tree is shown in Figure 2.28.

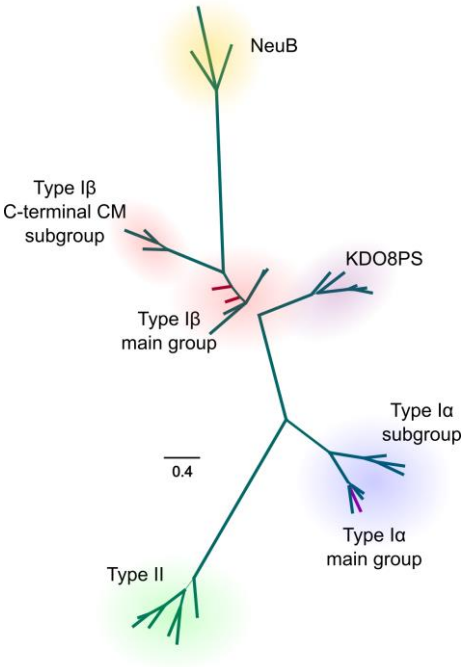


Figure 2.28. Unrooted Bayesian tree for a selection of DAH7PS, KDO8PS and NeuB sequences. The width of the branches is proportional to the posterior probability. Taxa are coloured by kingdom: dark cyan for Bacteria, dark red for Archaea and dark purple for Eukaryotes. The main clades of the tree are highlighted depending on which group they belong: red for type Iβ, blue for type Iα, green for type II, purple for KDO8PS and yellow for NeuB.

NeuB and KDO8PS do not form a monophyletic group, instead they both branch from type Iβ, but from different ancestors. NeuB split from the subgroup of type Iβ corresponding to C-terminal CM domain, whereas KDO8PS split from the main group of type Iβ. Moreover, the position of the subgroup of type Iβ and as well as the position of KDO8PS relative to the other major clades are fundamentally different in this tree than in the previous one. Two possible explanations stand: it might be due to the very small amount of sequences used to construct this tree that might incorrectly lead to the relative positioning of these two clades, or it is the improved alignment provided by the method that allowed this particular topology to be identified and new relationships to be refined. In order to correct potential flaws due to the limited number of taxa used, the method is currently being reapplied to a larger number of

sequences. However, this increased number of sequences lead to a very slow convergence of the chains and calculations are not yet finished. Nevertheless, two rooted versions of the tree shown in Figure 2.28 are presented in Figure 2.29, using NeuB as an outgroup (A), or midpoint rooting (B).

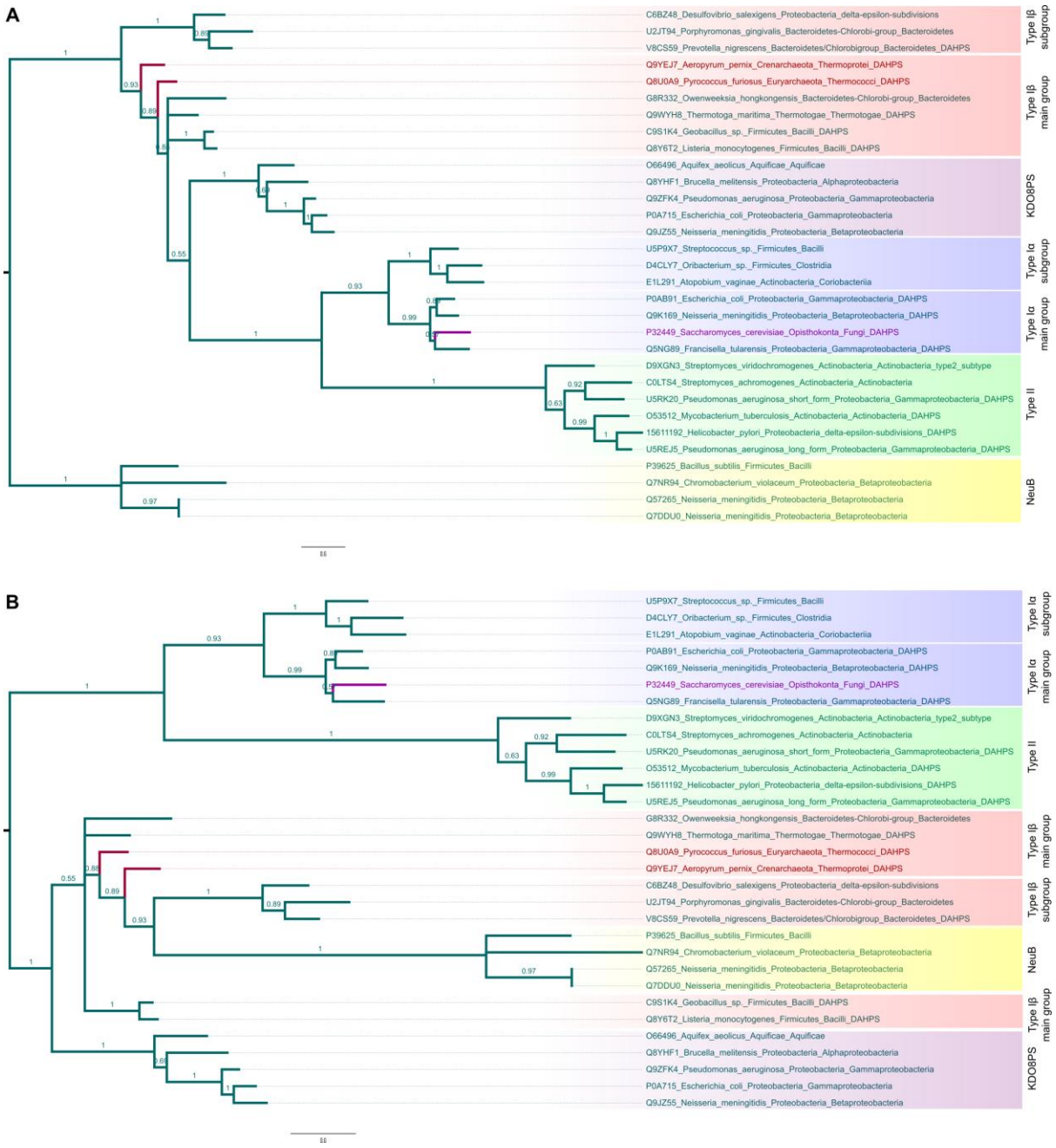


Figure 2.29. Bayesian trees for a selection of DAH7PS, KDO8PS and NeuB sequences rooted using NeuB as an outgroup (A) and midpoint rooted (B). Bayesian posterior probabilities are indicated on each branch.

Taxa are coloured by kingdom: dark cyan for Bacteria, dark red for Archaea and dark purple for Eukaryotes. The main clades of the tree are highlighted depending on which group they belong: red for type Iβ, blue for type Iα, green for type II, purple for KDO8PS and yellow for NeuB.

2.4 Discussion

Not only are type Iβ DAH7PS sequences represented among virtually all bacterial phyla, but they are proportionally more often found in extremophiles and deep branching Bacteria such as Firmicutes than in Bacteria that have diverged from a common ancestor more recently, such as Proteobacteria. Moreover, type Iβ is the only group found in Archaea and the only one not found in Eukaryotes. Eukaryotes are likely to have acquired type II and type Iα DAH7PS via horizontal gene transfer form Proteobacteria, in contrast with Archaea which might not be a distant lineage from common DAH7PS ancestor as they branch deeply within the type Iβ tree and are often unregulated. It is also in the type Iβ group that unregulated DAH7PS can be found and the allosteric regulatory domains correspond to N-terminal or C-terminal extensions with no insertions within the scaffold of the catalytic TIM barrel in the large majority of the cases. This type of allosteric regulation is primarily based on a mechanical blocking of the active sites of the tetramer by large regulatory domains (Chapter 4 and Chapter 5). In contrast, type Iα and type II rely on subtle regulation mechanisms [165,247] that utilise large domain or subdomain insertions. It is, therefore, reasonable to assume that evolution of allosteric regulation in type Iα and type II required more changes to the catalytic barrel to fine tune the regulations than for type Iβ. This abundance of evidence suggests that type Iβ is likely closer to the common ancestor of all DAH7PS and that other types of DAH7PS are likely to have evolved from an unregulated ancestor of type Iβ. It also seems reasonable to assume that allosteric regulation evolved later than the catalytic function. Based on these assumptions, type Iβ DAH7PS are

expected to have undergone fewer evolutionary changes than type II and type Iα DAH7PS and, therefore, this group is expected to be found closer to the root than type II and type Iα DAH7PS. This is observed in the trees shown in Figure 2.27 and Figure 2.29 (but not in trees a and b of Figure 2.26).

Allostery in type Iβ DAH7PS seems to have evolved via gene fusion, for example between a DAH7PS and a CM or ACT domain. Because ACT domains can share very little sequence similarity between one another and yet present the characteristic βαββαβ fold, little evolutionary pressure is expected on the sequence of the ACT domain and therefore evolution of ACT-domain regulated DAH7PS is primarily driven by the catalytic barrel, explaining why both ACT domain regulated and unregulated type Iβ DAH7PS appear very scattered and mixed together on the phylogenetic tree of type Iβ DAH7PS. In contrast, the CM domains have retained their catalytic activity leading to more evolutionary pressure on this subtype of DAH7PS, which are accounted for by their monophyletic character observed on the tree. The example of Streptococcus species having lost their N-terminal CM domain as well as the existence of clades containing both ACT-domain regulated and unregulated forms of DAH7PS indicate that loss of the regulatory domain might be frequent with retention of the catalytic barrel mutations gained prior to the loss of regulatory domain. Species that present the greatest degree of evolutionary changes have acquired, perhaps via loop extension, insertions within the catalytic barrels that might have fostered the apparition of more subtle regulation mechanism as observed in type Iα proteins. Indeed, the type Iβ DAH7PS found in *Nitrosococcus oceani* presents both an ACT domain as well as a longer $\beta_6\alpha_6$ loop, which is relatively similar to the N-terminal βαα extension and the $\beta_{6a}\beta_{6b}$ hairpin insertion of type Iα, or the $\beta_6\alpha_6$ insertion identified in the subgroup of type Iα.

It is known that organisms relying purely on type Iα for the synthesis of aromatic amino acids often express two or three isozymes, each being inhibited by a different aromatic amino acid. Because this is a rather inefficient way to regulate the flux within the shikimate pathway, it is possible that some organisms have evolved to express a single enzyme sensible to all three amino acids, giving rise to type II DAH7PS. Moreover, if the endogenous versus exogenous hypothesis [167] is correct and the presence of different isozyme permits a fine control of the production of aromatic amino acids depending on the environment, then it is not surprising that type II would have evolved to exploited the synergetic effect of ligand binding, in order to mimic the regulation observed with more than one enzyme. In addition, several organisms express both type II and type Iα DAH7PS, whereas none express both type II and type Iβ DAH7PS. It is thus possible that type II has evolved from type Iα or an ancestor of type Iα. After all the N-terminal domain of type Iα is only one helix shorter than type II, and the very long N-terminal extension found in the type Iα DAH7PS expressed by some organisms as *Metascardovia criceti* and *Bifidobacterium gallicum*, might be an illustration of how type II has started to evolve from type Iα. Furthermore, as discussed previously type Iα could have evolved from type Iβ or an ancestor of type Iβ, as the type Iβ DAH7PS that have experienced the most important changes, including insertions within the main barrel as it is the case in type Iα and type II, correspond to Proteobacteria and this phylum is, in proportion, more likely to express type II and type Iα. Perhaps these type Iβ Proteobacteria are relics of an evolutionary impasse toward the evolution process that yielded type Iα, or it might be a slightly more evolved lineage than the ancestor of type Iα. Similarly, it might not be a coincidence that a number of Bacilli species found in the subgroup of type Iα are also found in type Iβ. Overall, this points toward an evolution from type Iβ to type Iα to type II. The trees shown Figure 2.27 and Figure 2.29 are in good agreement with the this hypotheses as they depict an evolutionary scenario in which type II and type Iα share a common ancestor and this ancestor share a common ancestor

with type Iβ. Allostery in type Iα and type II might therefore have evolved by building upon the allosteric regulation domains of type Iβ via loop expansion and domain expension, facilitated by the low evolutionary pressure on the ACT domain found in type Iβ.

KDO8PS structures are extremely similar to unregulated type Iβ and are able to catalyse the condensation of PEP and E4P (as well as substrate A5P). In addition, the metal ion found in some KDO8PS, which is completely dispensable for the reaction it catalyses for some KDO8PS enzymes, might be a carryover from an common ancestor with metal-dependent DAH7PS [253]. Since KDO8PS is not allosterically regulated and catalyses a reaction similar to DAH7PS, it might have diverged from an DAH7PS ancestor during the evolution of the catalytic site but prior to the evolution of allostery in DAH7PS. It is therefore extremely unlikely that it represents a common ancestor of type Iα and type II, ruling out the tree presented in Figure 2.29A. This tree also misrepresents the evolution of NeuB. Indeed, NeuB has evolved to permit Gram-negative bacteria to mimic Eukaryotic cells, it is therefore a very recent enzyme, and DAH7PS, which is required for the biosynthesis of aromatic amino acids and KDO8PS, which is important for the cell wall of Gram-negative bacteria, represent essential enzymes that precede the apparition of Eukaryotes. Consequently, it is expected that NeuB has evolved recently. It is therefore possible that this enzyme only branches recently and presents a substantial amount of evolutionary modification form the ancestor of DAH7PS, which would be in disagreement with the tree presented in Figure 2.29A. On the other hand, an evolutionary relationship between the subgroup of type Iβ and NeuB could potentially make sense as both are dimeric proteins and both have C-terminal domain extension. Thus, although unlikely due to the major differences observed in the active site of NeuB when compared to DAH7PS and KDO8PS, evolution of NeuB from an ancestor of C-terminal CM DAH7PS remain a possibility

and so does the tree in Figure 2.29B, although this topology would need to be confirmed by including a greater number of sequences in the analysis.

Given the hypotheses formulated above on the grounds of experimental and evolutionary data, both the tree depicted in Figure 2.29B and the tree shown in Figure 2.27 represent possible scenarios that explain the evolution of DAH7PS with respect to NeuB and KDO8PS and the evolution of allostery within this family.

Chapter 3 – The allosteric regulation of type Iα DAH7PS from *Neisseria meningitidis*

3.1 Introduction

The DAH7PS enzyme expressed by the pathogen *Neisseria meningitidis*, which is responsible for meningitis, belongs to the type Iα class of DAH7PS. *N. meningitidis* DAH7PS is expressed as a single isozyme and is allosterically inhibited by Phe. Crystal structures of the inhibitor-free and Phe-bound enzyme do not indicate any major differences in the position of most backbone atoms. Molecular dynamics simulations of the apo and Phe-bound form of *N. meningitidis* DAH7PS do not reveal any large conformational changes even by conducting accelerated MD.

As this enzyme seems to be allosterically regulated in the absence of any major conformational changes (dynamic allostery), a number of the approaches, which have been outlined in Chapter 1, have been tried in order to understand how the binding at one site affects the catalytic activity at a distant site, and how the allosteric information is communicated from the allosteric site to the catalytic site. In particular, mutual information to identify correlated motions between alpha carbons [114] or dihedral angles [115], dynamical network analysis on alpha carbons [116] and residue centre of mass [118], identification of regions of high frustration [123], co-evolution analysis [254], statistical coupling analysis [158] can shed light

on dynamic allostery. However, when applied to *N. meningitidis* DAH7PS, none of these methods led to conclusive results about the way the allosteric signal is communicated from the allosteric site to the catalytic site.

In the absence of results from these methods in the case of *N. meningitidis* DAH7PS, I decided to develop and implement an original approach based on the analysis of pK_a variations. Ionisable residues do not present a constant pK_a in proteins, instead, their pK_a depends on the environment of each residue. The underlying premise of the method is that, even in the absence of major conformational changes, the environment of each ionisable residue involved in, or neighbouring a residue involved in the communication of the allosteric information, will change. Therefore, by monitoring the pK_a changes in the presence and absence of allosteric ligand, the mechanism of allosteric regulation can be deciphered.

This chapter consists in the article, recently published in the *Journal of the American Chemical Society*, presenting this method and its application to *N. meningitidis* DAH7PS:

Lang EJM, Heyes LC, Jameson GB, Parker EJ: **Calculated pKa Variations Expose Dynamic Allosteric Communication Networks**. *Journal of the American Chemical Society* 2016, **138**:2036-2045

Dr. Logan C. Heyes expressed and purified the enzyme, conducted the crystallographic experiments, including condition optimisation, crystallisation, soaking, data collection, refinement, deposition (PDB code: 4UC5) and analysis of the crystal structure. He identified possible allosteric communication pathways based on subtle atomic displacements in the apo and Phe-bound crystal structures. He performed pH-dependent isothermal titration calorimetry

(ITC) and pH-dependent kinetics and contributed to the writing of the manuscript. Prof. Geoffrey B. Jameson contributed to the writing of the manuscript. Prof. Emily J. Parker designed experiments and contributed to the writing of the manuscript.

I designed, prepared, performed and analysed the MD simulations. I invented the pK_a-based method, implemented it as a python programme and analysed the results. I contributed to the analysis of the ITC results and contributed to the writing of the manuscript.

3.2 Calculated pKa Variations Expose Dynamic Allosteric Communication Networks

The supporting information accompanying this article is available in Appendix B – Supporting information for Chapter 3.

Calculated pK_a Variations Expose Dynamic Allosteric Communication Networks

Eric J. M. Lang†, Logan C. Heyes†, Geoffrey B. Jameson§, and Emily J. Parker*‡

†Biomolecular Interaction Centre and Department of Chemistry, and ‡Maurice Wilkins Centre, Biomolecular Interaction Centre and Department of Chemistry, University of Canterbury, PO Box 4800, Christchurch 8140, New Zealand

§ Institute of Fundamental Sciences, Massey University, PO Box 11-222, Palmerston North 4422, New Zealand

J. Am. Chem. Soc., 2016, 138 (6), pp 2036–2045

DOI: 10.1021/jacs.5b13134

Publication Date (Web): January 21, 2016

Chapter 4 - The allosteric regulation of type Iβ DAH7PS from *Thermotoga maritima*

4.1 Introduction

The allosteric decoration of DAH7PS from the hyperthermophile *Thermotoga maritima* consists in an N-terminal ACT regulatory domain that binds allosteric inhibitor Tyr [180,255]. *T. maritima* DAH7PS exists as a tetramer in solution and has been crystallised in the absence (Protein Data Bank ID code: 1RZM) [179] and presence of Tyr (Protein Data Bank ID code: 3PG9) [180]. Each chain (Figure 4.1 A) is composed of 338 amino acids that can be divided into three distinct domains: the N-terminal regulatory ACT domain from residue 1 to residue 63 which consists in a typical βαββαβ structural motif [246], a flexible linker region (residues 64 to 82) and a core catalytic barrel from residue 83 to residue 338. A divalent metal ion is required for the catalytic activity of all DAH7PS enzymes; in the case of *T. maritima* DAH7PS, manganese is the metal coordinating ion for which the fastest reaction rates have been observed [256]. The catalytic TIM barrel possesses a classical $(\alpha/\beta)_8$ topology with an extended $\beta_2\alpha_2$ connecting loop (residues 127 to 143) covering the active site and being involved in the binding of E4P and PEP. The $\beta_8\alpha_8$ connecting loop (residues 299 to 315) also covers the active site, and is partially responsible for the binding of E4P and the metal ion, whereas the $\beta_4\alpha_4$

connecting loop contributes to the PEP binding site and is buried inside of the tetrameric structure.

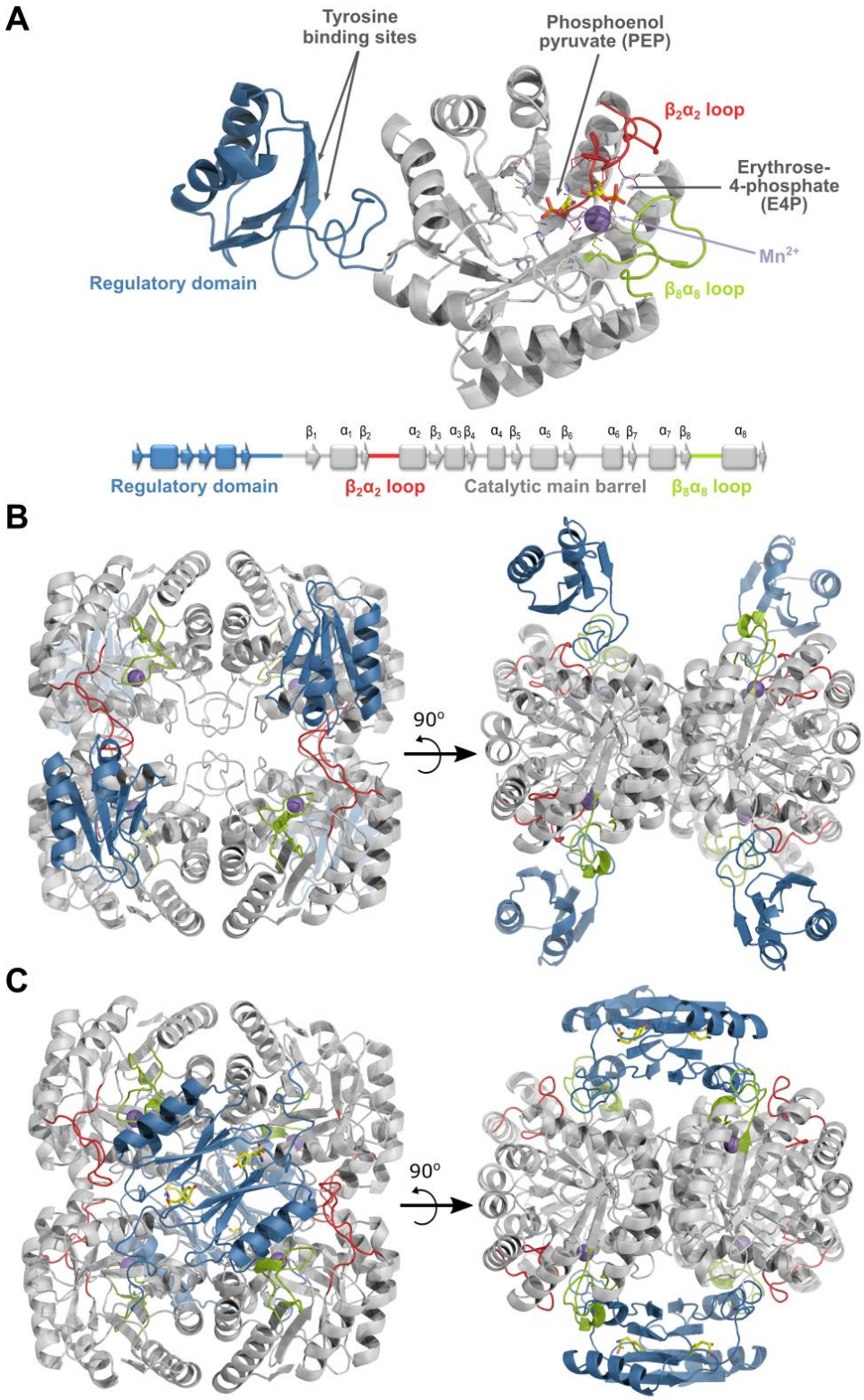


Figure 4.1 A. Structure of monomeric *T. maritima* DAH7PS. The regulatory domain is coloured in blue and the catalytic main barrel is coloured in light grey. Connecting loops $\beta_2\alpha_2$ and $\beta_8\alpha_8$ that are known to be catalytically

important are shown in red and green respectively. Substrates E4P and PEP are represented with fat sticks, metal ion Mn2+ is identified with a purple sphere and residues responsible for the binding of the substrates and metal ion are represented with thin sticks. B. Structure of tetrameric *T. maritima* DAH7PS in the absence and C. in the presence of allosteric inhibitor. The regulatory domains are coloured in blue and the catalytic main barrels are coloured in grey with $\beta_2\alpha_2$ and $\beta_8\alpha_8$ loops coloured in red and green respectively. Allosteric Tyr molecules bound to the regulatory domain interfaces are represented by yellow sticks.

Tetrameric *T. maritima* DAH7PS is mainly inhibited by Tyr, and to a lesser extent by Phe [257]. Comparison of the active and Tyr-inhibited crystal structures reveals remarkable differences in the position of the regulatory domains. Specifically, in the absence of allosteric ligand, the regulatory domains appear to stand in a vertical conformation (Figure 4.1B) whereas in the inhibited structure diagonally adjacent ACT-domains face each other in a horizontal fashion with the allosteric ligands bound at the interface (Figure 4.1C). This arrangement suggests that allosteric inhibition is achieved through a striking conformational change from an active broadly open form to an inhibited closed form.

However, it is not known in any detail how these ACT domain associations result in a decreased catalytic activity and what permits this association. Moreover, kinetic and SAXS experiments [180,257] reveal that solely relying on crystal structures is not sufficient for explaining the allosteric inhibition mechanism of this enzyme. For instance, it has been suggested based on SAXS data that in the absence of an allosteric ligand, *T. maritima* DAH7PS exists as an equilibrium between the open and closed forms and that the presence of Tyr shifts the equilibrium toward the closed form. Moreover, the inhibition profile of the enzyme in the presence of Tyr seems to plateau at around 15% of activity, demonstrating that even at high concentration of Tyr, the enzyme is still partially catalytically active.

In order to fully comprehend the allosteric regulation mechanism of *T. maritima* DAH7PS, a more in-depth understanding of the conformational dynamics in solution is required. Herein, the results obtained from all-atom molecular dynamics (MD) simulations performed to investigate the behaviour of the enzyme in solution are presented.

4.2 Methods

4.2.1 System preparation

Three crystal structures of *T. maritima* DAH7PS were used as starting points: the open form co-crystallized with E4P and PEP (PDB code: 1RZM), the closed form co-crystallized with Tyr (PDB Code: 3PG9) and the truncated mutant, which lacks the regulatory domains (PDB Code: 3PG8). Tetramers were prepared from the crystal structures. For the open form, E4P and PEP molecules were removed so the dynamic behaviour of the enzyme could be explored in the absence of ligands. Using COOT [258], cadmium ions in the active sites were replaced with manganese ions since catalytic activity and inhibition assays were performed in the presence of this metal [180]. Mn^{2+} parameters were added to the CHARMM force field parameter set based on the MM3 force field, as previously reported [165]. Missing residue Glu261 was added with COOT and the position of the side chain of one Met211 was changed to avoid steric clashes. For the closed form, two systems were prepared in order to explore the effect of the allosteric ligand on the closed system: one with Tyr bound to the allosteric sites and one without any allosteric ligand. Manganese ions were added to the metal-free crystallographic structure in COOT based on the metal coordinates of the open form and missing atoms side chains were added with COOT. For the truncated mutant, missing residues were reconstructed with MODELLER [259] using the automodel routine. In order to correctly

model the position of the missing metal binding residues, metal binding residues in 1RZM, as well as their direct neighbours, were used by homology. The non-natural amino acid found at position 102 of 3PG8 was also modelled as a Cys. A total of 100 models were created with MODELLER using a thorough optimization procedure that is repeated twice per model (setting `automodel.library_schedule` to `autosched.slow`, `automodel.max_var_iterations` to 300 and `automodel.md_level` to `refive.very_slow`) . Only the missing residues and the residues missing atoms were allowed to move, the others being constrained to their crystallographic position. The best model was chosen based on both the best DOPE and MolPDF scores as well as on the overall best DOPE score per residue profile The best model was also checked visually to ensure a realistic positioning of the missing residues based on the position they adopt in the known crystal structure. The protonation state of the ionisable residues was calculated for both the open and closed form using MCCE [260] for His residues and PROPKA 3.1 [261] for all the others, averaging the results per chain. In the case of differences between open and closed forms in the prediction, the protonation state determined for the closed form was kept and apply to the open, closed and truncated systems. Each system was solvated in an explicit TIP3 water box of 109 Å x 97 Å x 99 Å, which allows for the movements of the regulatory domains during the MD simulations without any risk of interactions with the periodic images. The net charge of each system was neutralised with Na^+ and Cl^- ions added with a minimum distance of 5 Å from the enzyme and from each other.

4.2.2 Classical MD simulations

All MD simulations were carried out with NAMD [89] using CHARMM all-atom parameter set 22 with the CMAP correction [82,262]. The simulations were run on the high-performance computing service facility at the University of Canterbury. The particle mesh

Ewald (PME) method was used to calculate electrostatic interactions and Van der Waals cut-off was set up at 12 Å. The simulations were performed at 363 K with the Langevin thermostat and with the Nosé-Hoover Langevin piston used to keep a constant pressure of 1 atm during equilibration and production runs. The simulations were conducted with a 2 fs time steps with the SHAKE algorithm used to constrain all bonds involving hydrogen atoms. Using the conjugate gradient energy minimization method, each solvated protein was minimized for a total of 100000 steps, first by allowing only the hydrogen atoms to move and keeping all the other atoms fixed (10000 steps), then water molecules and ions were released (10000 steps), followed by protein side chains (30000 steps) and finally, all atoms were released (50000 steps). Each system was gradually heated from 0 to 49 K, increasing the temperature by 1 K every 200 steps with all heavy protein atoms restrained in their crystallographic position with a 1 kcal/mol/Å^2 harmonic potential and then from 50 to 369 K, increasing the temperature by 1 K every 200 steps with all backbone protein atoms restrained in their crystallographic position with a 1 kcal/mol/Å^2 harmonic potential. Equilibration using isobaric-isothermal conditions for 27200 steps with all backbone protein atoms restrained in their crystallographic position with a 0.5 kcal/mol/Å^2 harmonic potential was performed and followed by equilibration using isobaric-isothermal conditions for 300000 steps with all atoms free to move. The MD simulations were started with a different velocity distribution each time and the trajectories were written out every 100 ps. Two simulations were run for the open form, one for 635 ns (Open_run1) and the other for 335 ns (Open_run2). An additional simulation of 335 ns at 300 K was performed (Open_300K). For the closed form, 2 simulations of 335 ns with Tyr bound were run (Closed_run1 and Closed_run2) as well as one simulation performed in the absence of Tyr (Closed_noTyr). Finally, one simulation of 335 ns was run for the truncated mutant (Truncated).

4.2.3 Biased MD simulations

Starting with the equilibrated open and closed (without Tyr) systems, targeted molecular dynamics (TMD) simulations were run from the open to the closed form and from the closed to the open form using a force constant 600 kcal/mol/Å^2 applied to the alpha carbon of residues 1 to 80. The TMD simulations were run over 5,000,000 steps (equivalent to 10 ns of classical MD) using the crystal structure of each system as a reference and setting the RMSD of the final TMD steps to 0.2 Å.

A snapshot from the trajectory of the simulation Open run 1 was taken at 31.7 ns to serve as a starting point to run two accelerated MD (aMD) simulations. After extensive testing, it was found that the boost potential parameters *E(dih)* and *α(dih)* defined using the method of Tikhonova and co-workers [263] gave the best results. In this approach the boost potential is exclusively applied to the dihedral angles following $E(dih) = < V_0(dih) > + < V_0(dih) > \times c$, and $\alpha(dih) = E(dih)/5$, where $<V_0(dih)>$ is the average dihedral potential energy and *c* is a constant with a value comprised between 0.2 and 0.5 in the original paper. In this work we used the first 31.7 ns of simulation to assess the average dihedral potential energy and the best results were obtained for c = 0.4 (moderate acceleration) and c = 0.8 (aggressive acceleration).

4.2.4 Trajectory analysis

Trajectories were aligned with respect to the TIM barrels to remove undesired rotational and translational motions of the enzyme. Root mean square deviations (RMSD), root mean square fluctuations (RMSF), radius of gyrations (Rg) and dihedral angles ϕ and ψ were calculated using Gromacs 4.5 [264]. Hydrogen bonds, distances, angles, solvent-accessible

surface area (SASA) and active site occlusion were calculated in VMD [265] using purpose made Tcl scripts. The free energy surface was calculated by projecting of the trajectory onto the two coordinates of interest and dividing the space with a 100 × 100 grid. The free energy for each bin is then given by $\Delta G = - RT \ln(P_i/(1-P_i))$ with $P_i = N_i/N$, where N is the total number of frames and N_i is the number of frames in each bin. The energy of the unsampled regions is not defined. Solvent accessible cavities leading to the active sites were identified using KVFinder [266] using a probe 'in' radius of 1.4 Å, a probe 'out' radius of 6.0 Å and a volume filter of 5.0 $Å^3$ on a selection of snapshots from the MD simulations. Structures and ligands for docking experiments were prepared with AutoDock Tools [267] and the docking was performed with AutoDock Vina [268].

4.3 Results

4.3.1 Large amplitude movements of the regulatory domains are observed for the open form

Classical MD simulations of *T. maritima* DAH7PS in its open state at the upper growth limit temperature of the organism 363 K [269] (Open_run1 and Open_run2) present a rapid and important increase in the RMSD with respect to the open form crystal structure (Figure 4.2), exceeding 8 Å after 50 ns simulated. The RMSD traces do not plateau after the initial burst but instead oscillate between approximately 7.5 and 11.5 Å. Pursuing the simulation for longer to reach 635 ns does not help to attain a plateau, but instead further increases the deviation from the initial structure. This behaviour suggests that the protein experiences large conformational changes without reaching an energy minimum deep enough to remain in a stable conformation. This is confirmed by visual inspection of the MD trajectories (Movie 4.1

and Movie 4.2 – Appendix A – Supplementary movies), which show that the regulatory domains experience major conformational changes due to the flexibility of the linker region, while the catalytic domains do not present any significant alteration of their conformation. Each ACT domain samples a large portion of the accessible conformational space and only temporally remains relatively immobile when it engages in interactions with the adjacent ACT domain or with the catalytic barrels. Unsurprisingly, because of the degree of freedom of the regulatory domains imparted by the flexible linker region, classical MD is not sufficient to sample the fully closed form starting from the open form crystal structure.

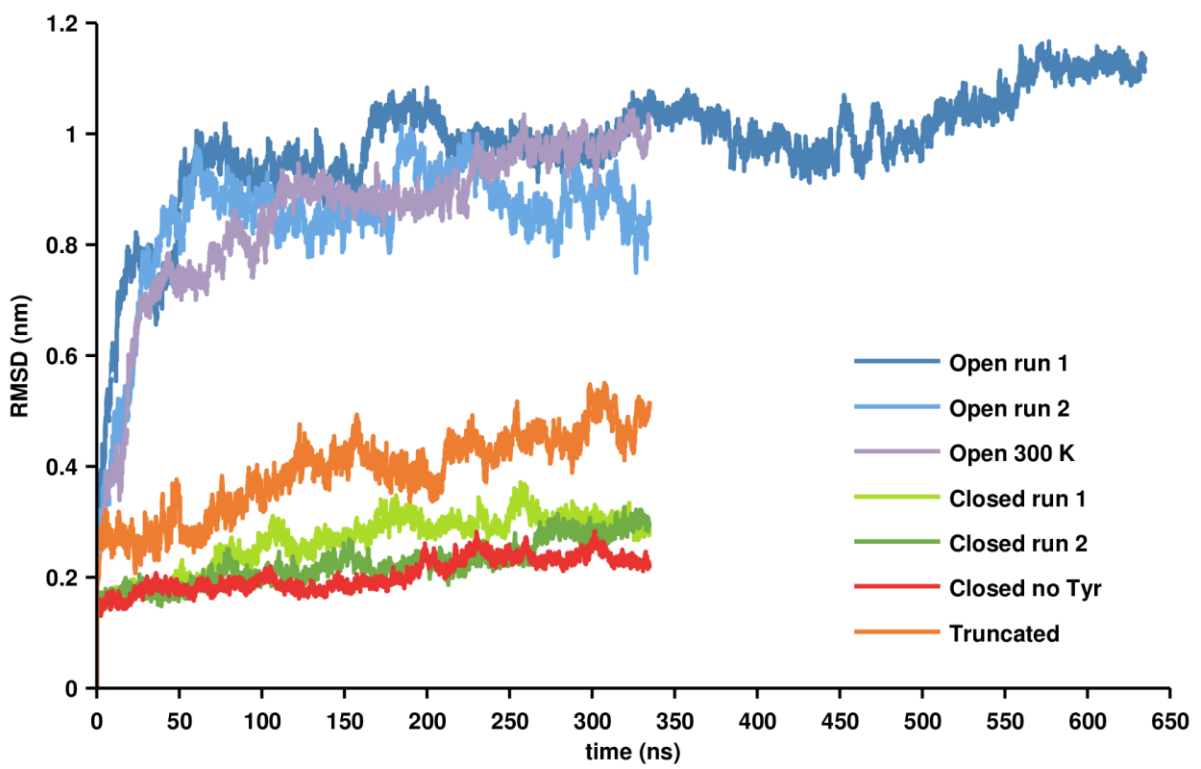


Figure 4.2. Alpha carbon RMSD traces of the MD trajectories of tetrameric *T. maritima* DAH7PS. For the open form simulations (Open_run1 shown in dark blue, Open_run2 in light blue and Open_300K in purple) the RMSD is calculated with respect to the open form tetrameric crystal structure (PDB ID: 1RZM), while for the closed form (Closed_run1 shown in light green, Closed_run2 in dark green and Closed_noTyr in red) and truncated mutants (Truncated shown in orange), the RMSD is calculated with respect to the tetrameric closed crystal structure (PDB ID: 3PG9) and truncated tetrameric crystal structure (PDB ID: 3PG8) respectively.

Interestingly the broadly open conformation observed in the crystal structure does not last for more than a few nanoseconds, almost immediately the regulatory domains deviate from their original position and start exploring the surrounding area. This seems to indicate that the conformation observed in the crystal structure is not stable in solution but might exist only because of the crystal packing in solid phase. Closer inspection of the crystal structure reveals that each ACT domain interacts with two symmetry mates that are not part of the tetramer: one interacts via its main barrel, whereas the other symmetry mate interacts via its ACT domain (Figure 4.3). Several non-bonding interactions can be identified between the ACT domain and these two symmetry mates. The result is close-packing with interlaced ACT domains, which maintain the broadly open conformation observed in the crystal structure. In a diluted solution, when these inter-tetramer interactions do not exist anymore, the ACT domains rapidly leave the now-unstable broadly open conformation for a more relaxed conformation and start exploring the surrounding space.

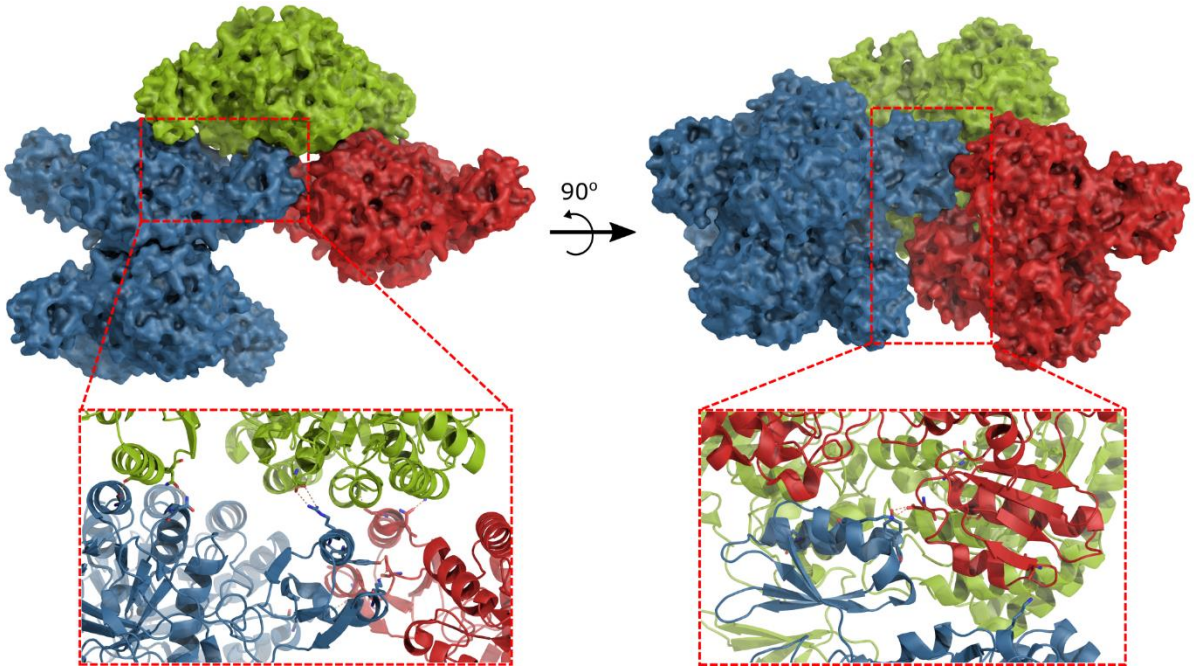


Figure 4.3. Surface representation of symmetry mates in crystal structure 1RZM. One of the ACT domain of tetrameric *T. maritima* DAH7PS (blue) is shown to interact with the ACT domain of an adjacent symmetry mate (red). Both ACT domains interact with the main barrel of a third symmetry mate (green). Zooms show the proteins as a cartoon and the main interacting residues are represented with sticks and the interactions with orange dotted lines.

In order to investigate the effect of temperature on the dynamics of the regulatory domains and confirm that the instability of the broadly open form in solution is not an artefact of the simulated temperature, a simulation was performed at 300 K (Open_300K). Although the initial increase in RMSD is not as steep as at higher temperature, the RMSD rapidly reach the same range of values as observed at 363K. Visualisation of the trajectories confirms the existence of large regulatory domain movements leading to interaction with the adjacent ACT domain or with the main barrels, even though the amplitude of these movements is slightly less important at 300 K (Movie 4.3 and Movie 4.4 – Appendix A – Supplementary movies).

In order to try to simulate the closing of the regulatory domains, two accelerated MD simulations were performed. During aMD simulations, a boost potential is applied if the energy falls below a chosen threshold. This boost potential enables to flatten the energy landscape by reducing the depth of the energy minima and thus lowering the height of the energy barriers to cross. The challenge of aMD is to find the appropriate acceleration parameters for the studied system, i.e. the acceleration factor α which determines the shape of the modified potential and the threshold energy E which controls the portion of the potential surface affected by the boost. Here the boost potential was only applied to the dihedral angles using a previously reported method (see Methods) and used two degrees of acceleration. In order for aMD to have an improved sampling over the classical MD runs, the aMD runs were also run at 363 K to take advantage of the increased sampling provided by the high temperature. However, the

combination of high temperature and boost potential led to the partial unfolding of the enzyme if the simulations were carried out for too long. As a result, the aMD simulations could only be run for 175 ns in order to limit to a minimum the unfolding of the secondary structure features of the enzyme. This period of time was insufficient to observe any closing events and the exploration of the conformational landscape over this period is identical to what is observed for longer unbiased simulations. Consequently, the results of these aMD simulations are not shown.

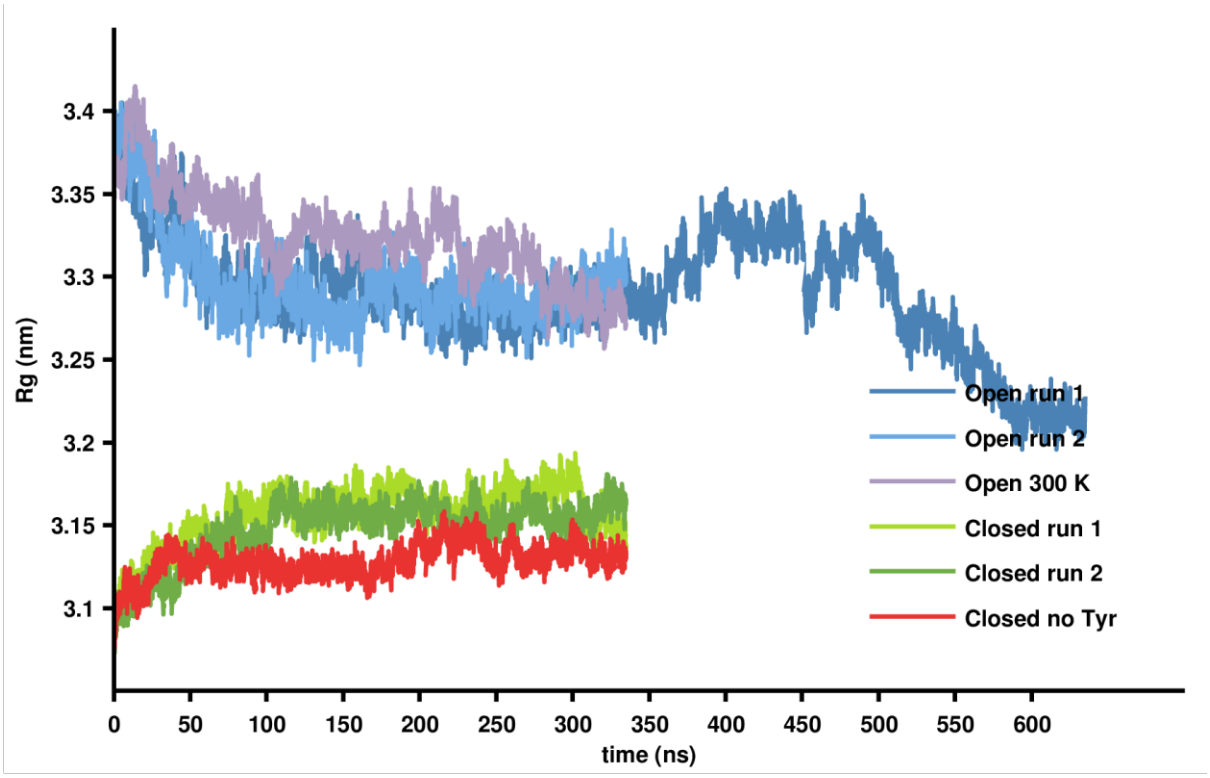


Figure 4.4. Radius of gyration of the tetramer as a function of simulated time. Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

To quantify the effect of the regulatory domain movements on the tetrameric structure, the radius of gyration *R*g of the tetramer was monitored over the course of the simulations as a measure of the compactness of the protein (Figure 4.4). For the open form trajectories, the *R*g decreases rapidly from its initial value and then tends to decrease at a lower rate, sometimes

plateauing, sometimes increasing for some time. This is in line with the unrestrained movements of the regulatory domains observed at the beginning of the simulation and then the creation and breaking of interactions with an adjacent regulatory domain or with the catalytic barrels. Overall the motions of the regulatory domains exploring their available conformational space correspond to closing motions which lead to an increased compactness of the tetramer in solution that can even approach the level of compactness observed for the closed form.

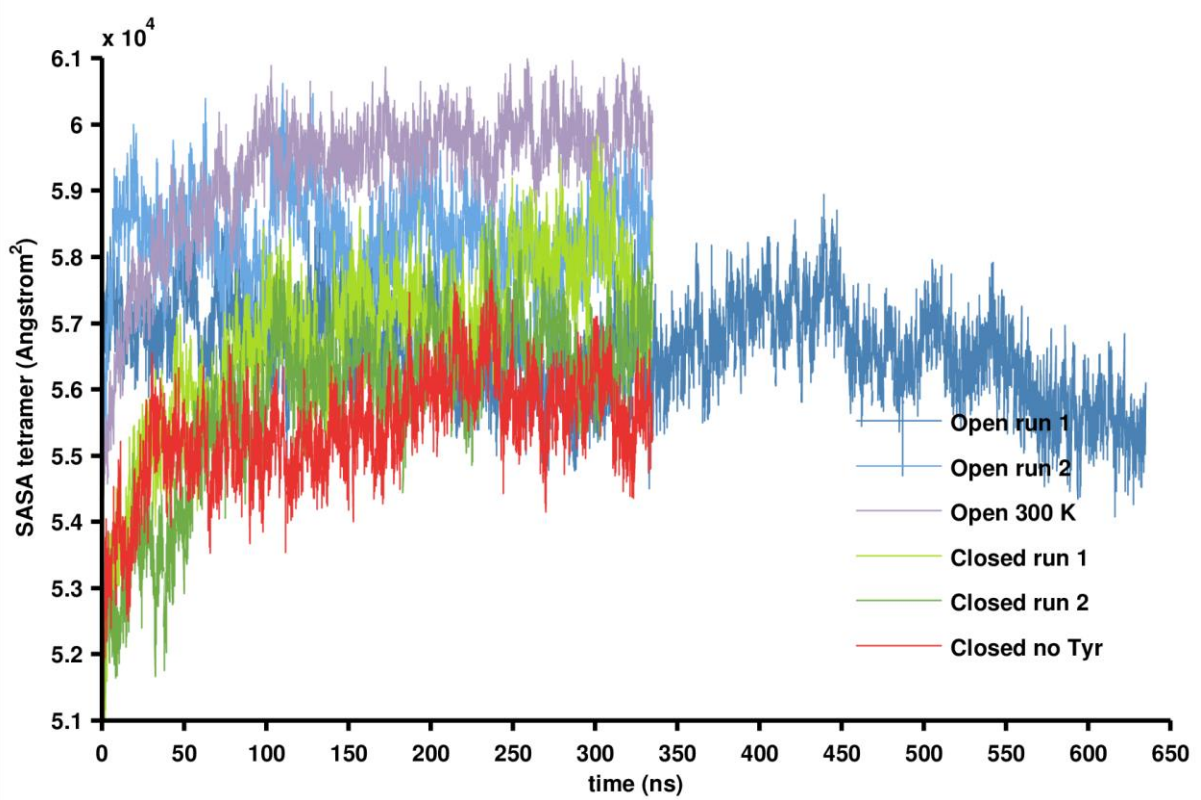


Figure 4.5. Evolution of the total solvent accessible surface area of the tetramer over the course of the simulations. Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

In order to understand the effect of the increased compactness on the open form, the solvent-accessible surface area (SASA) of the tetramer was calculated for each frame of the trajectories as shown in Figure 4.5. For the open form simulations, after an initial burst

corresponding to the relaxation of the systems, the SASA oscillates and can decrease to reach a lower value than for the X-ray crystal structure. This again agrees perfectly with the fact that the regulatory domains interact with each other or with the catalytic barrel, leading to a reduction of the SASA. As these closing motions are less pronounced at 300 K than they are at 363 K, the effect on SASA is significantly less pronounced.

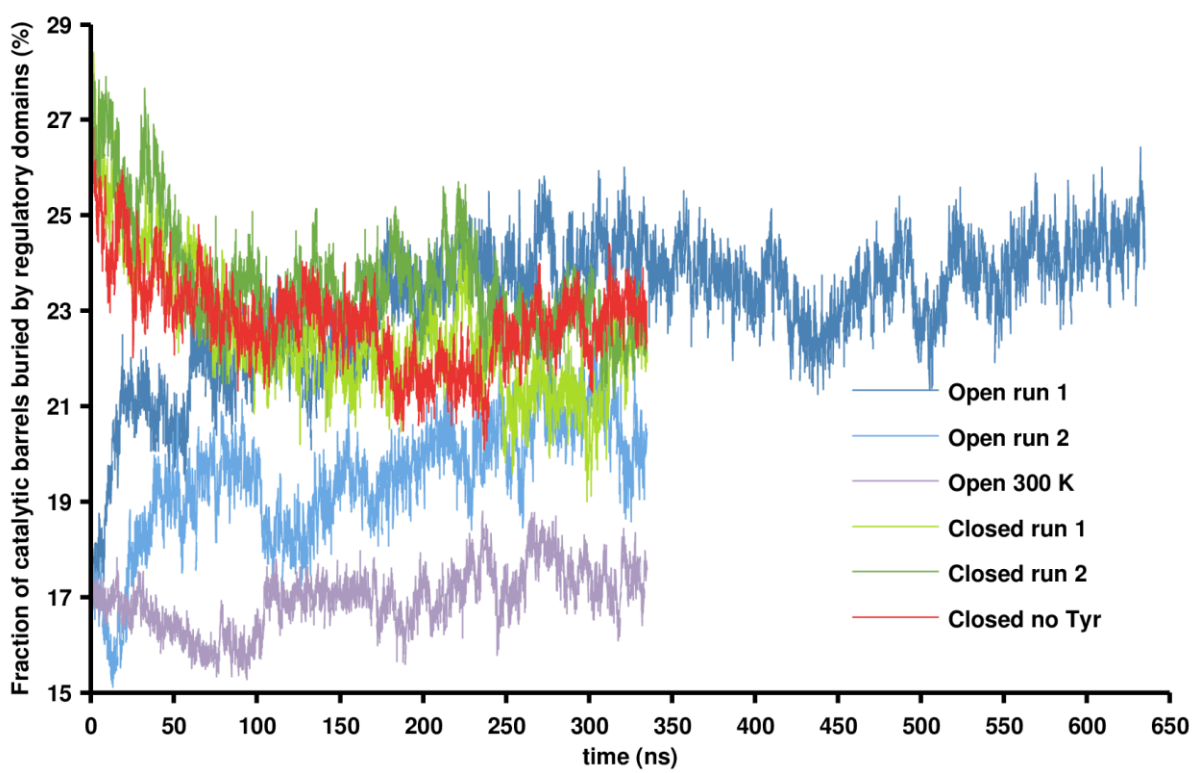


Figure 4.6. Fraction of the catalytic barrel buried by the regulatory domains in the tetramer as a function of simulated time. Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

SASA calculations additionally show that the fraction of the catalytic barrel buried by the ACT domains and the linker region tends to increase over the course of the open form simulations (Figure 4.6), which is consistent with the movements of the regulatory domains observed for these simulations. Large discrepancies are however observed across these simulations, with a fraction of buried catalytic barrel remaining around this initial value of 17%

in the case of Open_300K, it increases to up to 21% in the case of Open_run2 and 25% in the case of Open_run1. Such differences are not observed when comparing the RMSD traces (Figure 4.2) or *R*g evolution (Figure 4.4). This highlights the fact that the motion of the regulatory domains in the open form simulations are not necessarily accompanied by direct interactions with the catalytic barrels and instead, an ACT domain might simply interact with the adjacent ACT domain or can move toward the catalytic barrel without actually entering into contact with it, resulting in decrease in radius of gyration but a fraction of buried catalytic barrel remaining constant.

4.3.2 The closed form remains very stable throughout the trajectory even in the absence of allosteric ligand

In contrast with the MD simulations of the open form, the closed form remains very stable throughout the simulations, with the RMSD increasing slowly but constantly over the course of the simulations, but without going further than 2 to 3 Å from the closed form crystal structure (Figure 4.2). No opening of the regulatory domains was observed over the course of the simulations, even in the absence of Tyr in the allosteric binding site, suggesting that the closed state is a very stable conformation. Even a Tyr unbinding event observed in one trajectory (Closed_run1) did not trigger the opening of the regulatory domains (Movie 4.5 and Movie 4.6 – Appendix A – Supplementary movies), although its RMSD is more elevated when compared to the other two closed simulations (Closed_run2 and Closed_noTyr).

For the closed form trajectories in the presence of Tyr, the *R*g increases slowly from its initial value until reaching a plateau after about 125 ns and in the absence of Tyr, the *R*g stabilised sooner. In both cases, although the initial increase during the first 10-20 ns can

possibly be explained by a full relaxation of a system, which would be in line with the initial increase in RMSD observed for these trajectories during the same amount of time (Figure 4.2), the subsequent increase that takes more than 100 ns before to stabilised, cannot be solely explained by a relaxation of the system and instead other mechanisms are likely to be at play.

Globally the SASA of the entire protein increases gradually in the closed form and although the initial SASA is lower than for the open form, it can exceed the SASA of the open form simulations at 363 K.

4.3.3 Exploring the conformational energy landscape accessible to T. maritima DAH7PS

To better visualise the differences between the open form and closed form simulations in terms of sampled conformational space, their complete trajectories were projected onto the conformational space defined by the RMSD from open form versus RMSD from closed form as presented Figure 4.7.

The closed form simulations deviate by more than 15 Å from the open crystal structure, exceeding the RMSD difference between the closed and open forms of 14.2 Å. This is explained by the fact that the deviation from the closed crystal structure is not associated with an opening of the regulatory domains, thus, the more they deviate from the closed form, the further apart they are from the open form. Because of the absence of major global domain motions observed during the closed form simulations, their projection onto the RMSD space remains extremely localised, with large overlapping portions between the three closed form trajectories.

As observed in Figure 4.2, the open form simulations deviate rapidly from the open form crystal structure leading to the trail of sporadic conformations presenting an RMSD between 2 to 7 Å from the open form crystal structure (Figure 4.7). The bulk of the trajectories occupies a larger area with less overlap than what is observed for the closed trajectories. In addition, because of the partial closing observed during these simulations, an increase in RMSD from the open form roughly corresponds to a decrease in RMSD from the closed form. Although larger than for the closed form simulations, the portion of the conformational space sampled during the open form simulations represents only a small area of the vast conformational space that describes the opening and closing of the four regulatory domains and remains distant from the closed crystal structure by more than 10 Å.

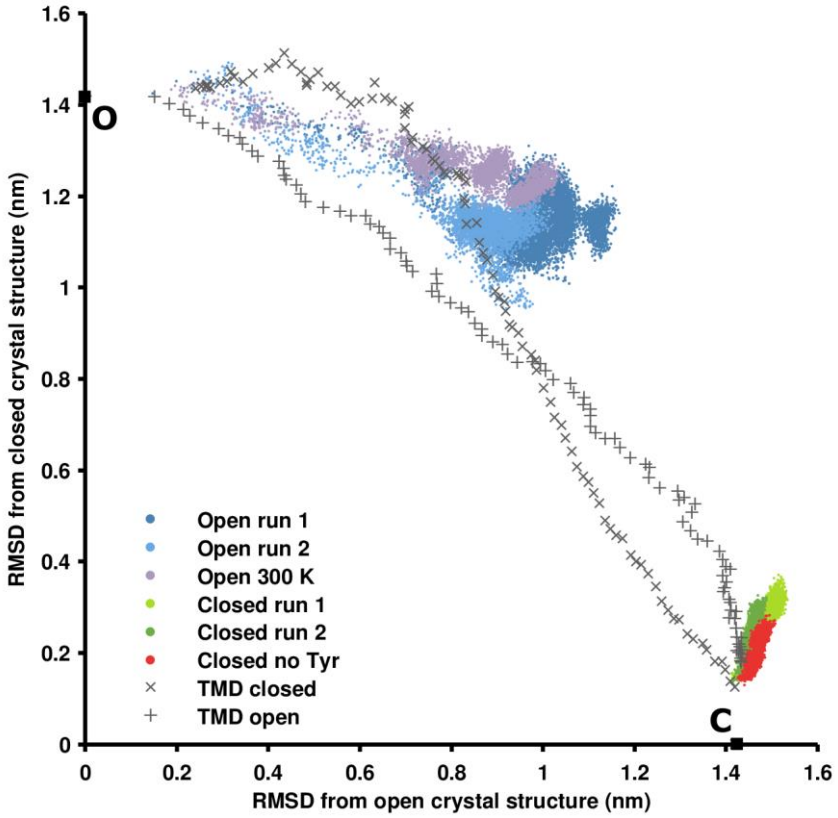


Figure 4.7. Projection of the trajectories onto the alpha carbon RMSD space formed by the RMSD calculated from the tetrameric open crystal structure (x-axis) and from the tetrameric closed crystal

structure (y-axis). Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red. The TMD trajectories are represented with grey crosses and the x-ray crystal structure of the open (O) and closed (C) forms are identified with squares.

To further investigate the amplitude of the movements required by the regulatory domains to reach a fully open or a fully closed form starting from a fully closed and fully open state respectively, a targeted molecular dynamics (TMD) simulation was run, using the crystal structures of the open and closed tetrameric forms as reference points and driving the transitions between these two final states by applying a harmonic potential force, based on the RMSD from the reference state, to the regulatory domain and linker region alpha carbons.

Projection of the open to closed (TMD_open) and closed to open (TMD_closed) TMD trajectories along the two RMSD coordinates better helps to visualise the amplitude of the conformational changes involved in the transition between open and closed forms which present an RMSD difference of 14 Å. The two trajectories loosely follow the line formed by the RMSD between the two crystal structures. Characteristic of TMD simulations, large amplitude global motions, which rapidly decrease the RMSD from the reference structure occur prior to small local changes. As a result, the RMSD plot first follows a quasi-rectilinear trace corresponding to the global changes. This region is comprised between 14 and 5 Å (RMSD from closed crystal structure) for the open to close simulation, and between 14 and 8 Å (RMSD from open crystal structure) for the close to open simulation. The traces then gain in curvature with a dramatic slope decrease, corresponding to local rearrangements. This behaviour is likely to be unrealistic as it is expected that both global and local changes would occur concomitantly. It is, however, reassuring that the region of local changes roughly follows the region sampled by classical MD. Therefore, although the resulting trajectory is not at equilibrium, it permits

the concomitant closing and opening of the regulatory domains to be simulated, otherwise impossible by conventional MD.

As noted above, visual inspection of the open form trajectories indicates that the regulatory domains occasionally and individually sample conformations that resemble those sampled in the closed form simulations. In order to better study this phenomenon, each trajectory of the tetramer was decomposed into four trajectories, one for each monomeric unit, which can be seen as four independent trajectories for the purpose of considering the sampling of the conformational space accessible to a single regulatory domain. This leads to 1.34 μs of monomeric unit dynamics for each system except for Open_run1, for which 2.54 μs of monomeric unit dynamics are available.

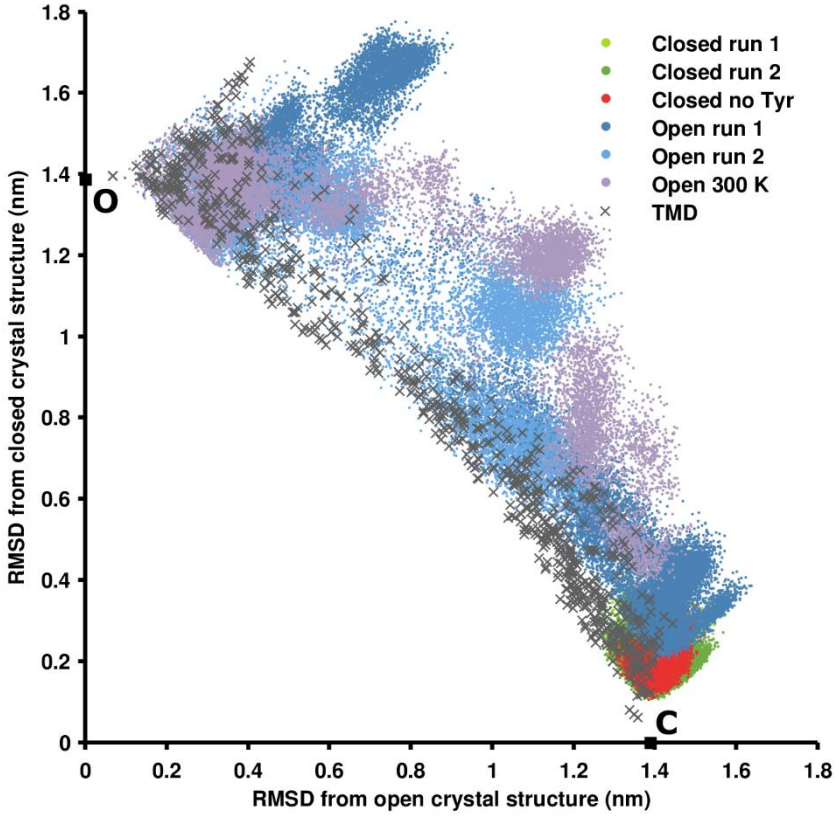


Figure 4.8. Projection of the trajectories onto the alpha carbon RMSD space formed by the RMSD calculated from the monomeric open crystal structure (x-axis) and from the monomeric closed crystal

structure (y-axis). Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red. The TMD trajectories are represented with grey crosses and the x-ray crystal structure of a single chain of the open (O) and closed (C) forms are identified with squares.

Projection of the monomeric unit trajectories onto a conformational space formed by the RMSD from the open or from the closed monomeric unit crystal structure is shown in Figure 4.8. Contrary to the conformational space of the tetramer, a more exhaustive sampling of the conformational space is achieved. Of particular interest, some regions sampled during the open form simulations overlap with regions of the conformational space sampled during the closed form simulations. Although the conformations obtained from the open state trajectories still present at best an RMSD of 2 Å from the closed crystal structure, more relaxed closed form conformations are sampled during the open form simulations. Therefore, it indicates that, although simulating the transition from a fully open tetrameric conformation to a fully closed one remains impractical using classical MD simulations, simulation of the closing of individual regulatory domains is amenable within the timescale accessible by conventional MD.

Moreover, because this conformational space is extensively explored during the simulations (combining all the simulations at 363 K corresponds to 7.9 μs of monomeric unit dynamics), a population-based relative free energy map of the conformational landscape of the monomeric unit in the tetramer (since the conformations sampled are dependent on the interactions with the adjacent regulatory domain or with the catalytic barrels) at 363 K can be calculated with fair accuracy. Here the energy map shown in Figure 4.9 was constructed along two coordinates representative of the closing/opening motions of individual domain: the radius of gyration of the monomeric unit and the angle formed by Gly33-Pro68-Lys231, which describes reasonably well the pendulum-like motion of the regulatory domain from open to

close and vice versa (Figure 4.10). For the construction of the energy map, the first 35 ns of each tetramer simulation were discarded to allow for extensive equilibration of each system and to exclude the initial RMSD burst observed at the beginning of the simulations (Figure 4.2).

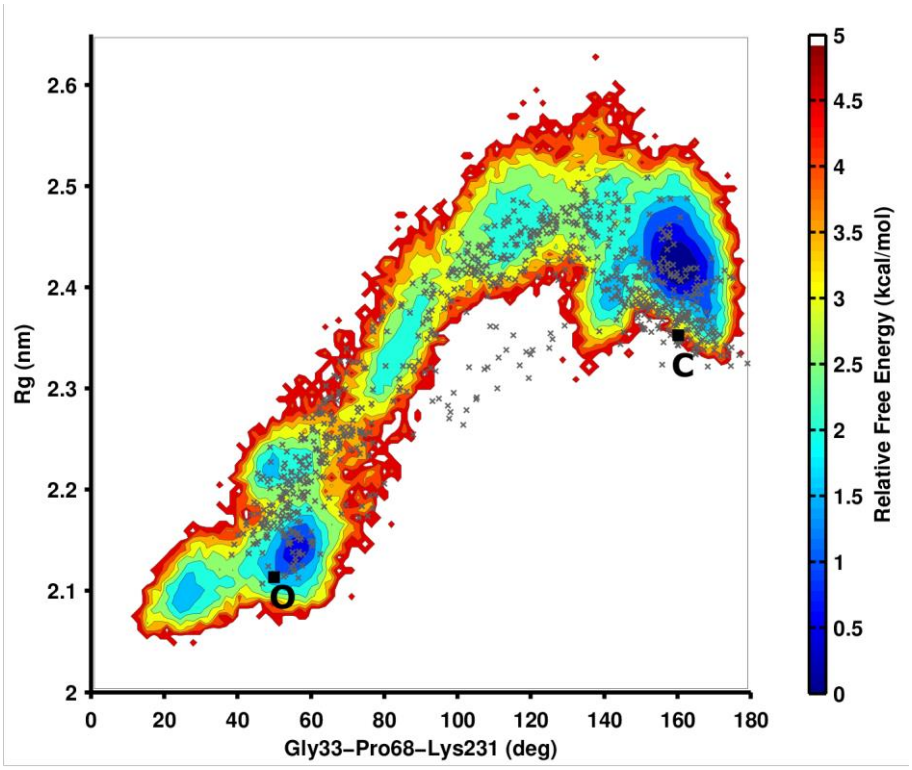


Figure 4.9. Relative free energy landscape formed by the Gly33-Pro68-Lys231 angle (x-axis) and the radius of gyration (y-axis). The map was calculated from the population obtained during the simulations at 363 K. The TMD trajectories are represented with grey crosses and the x-ray crystal structure of a single chain of the open (O) and closed (C) forms are identified with squares.

It is clear from this map that the X-ray crystal structures of the open and closed forms do not correspond to the calculated energy minima in solution but instead appear to be at the edge of the two deepest energy wells, centred on 54.9°, 2.14 nm and 159.3°, 2.42 nm, corresponding to the relaxed open and relaxed closed forms in solution respectively. Unsurprisingly the *R*g of the crystal structure monomeric units is lower than for the relaxed structure, likely due to the crystal packing effect. Overall the energy landscape presents an upside-down J-shape, which

is explained by the fact that the highest radius of gyration does not correspond to the largest angle but instead occurs before the relaxed closed state.

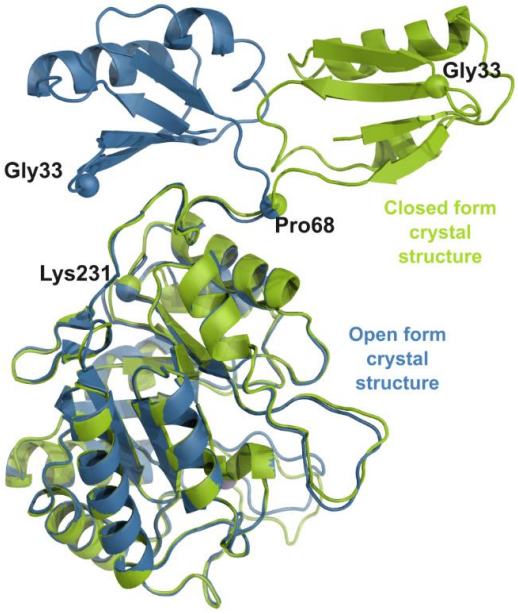


Figure 4.10. Structure of the X-ray crystal structures of the monomeric unit in its open form (blue) and closed form (green). Alpha carbon of Gly33, Pro68 and Lys231 used to calculate the angle representing the regulatory domain motions are represented by spheres.

The energy well corresponding to the relaxed closed state is only 0.6 kcal/mol deeper than the relaxed open form but the region covered larger. However, because no opening of the regulatory domains in the closed state was sampled, this energy map is flawed toward the closing of the regulatory domains, and the difficulty of opening the regulatory domains suggests that the energy well corresponding to the relaxed closed conformation is certainly much deeper than what is calculated here. The highest energy barrier is therefore expected to be encountered by a closed conformation willing to escape the closed conformation energy well. Nevertheless, the rest of this energy landscape is not expected to be fundamentally affected by additional sampling and it provides useful information regarding the closing mechanisms of the regulatory domain.

To reach a relaxed closed conformation starting from an open one, crossing of a barrier of about 3 kcal/mol with respect to the open form energy well is required. However, this barrier crossing does not occur in one step, and instead intermediate low energy states about 1.5 kcal/mol higher than the energy of the open state are spread along the path. Moreover, these shallow energy basinsmake the exploration of the conformational space between the open and close state facile. This region of the energy landscape accounts for the large number of conformations sampled with relative ease by the regulatory domains in the open form simulations.

With the exception of a few conformations, the TMD trajectories follow reasonably well the path between the two end states along these two coordinates and seem to sample both intermediate low energy states and higher energy regions. The trajectories also present a very good overlap with the region of the space characterised by the two RMSD coordinates, sampled by the classical MD simulations (Figure 4.9).

4.3.4 Three residues appear to be responsible for the flexibility of the linker region

As the TMD simulations provide as a whole a good approximation of single regulatory domain closing and opening mechanisms, they can inform on the role played by the residues composing the linker region during this process, in particular, those responsible for the opening and closing movements. Indeed, crystal structures and classical MD simulations reveal that the numerous degrees of freedom of the ACT domains are permitted by the flexibility of the linker region. In order to investigate the contribution of each residue, dihedral angles ϕ and ψ for

residues 63 (corresponding to the last residue of the ACT domain) to residue 70 were monitored during the TMD simulations.

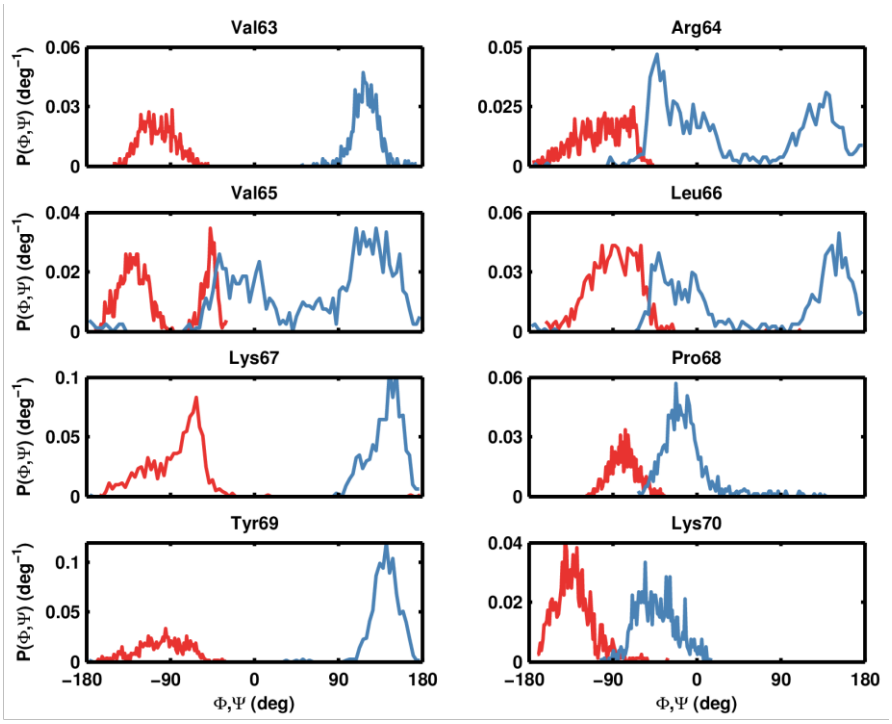


Figure 4.11. Frequency distribution plots of dihedral angles for residues 63 to 70 from the TMD simulations. ϕ dihedrals are coloured in red and ψ dihedrals in blue.

Linker residues after residue 70 show only marginal changes in their dihedral angles and their lack of involvement in the flexibility of the linker region is confirmed by their low RMSF values (see below). Plots of the frequency distribution of these dihedral angles of interest are shown in Figure 4.11. Three residues, Arg64, Val65 and Leu66 (Figure 4.12) stand out from the others by presenting an unusually broad distribution of their ψ angles that span more than 170° while their ϕ angles range from -180° to more than -45°. The flexibility of these residues seems to be essential for the free motions of the regulatory domains, and they appear to be the key to the closing mechanism. Of these three residues, Val65 presents the greatest dihedral angles changes and because of its position in between the other two, it is likely to be the most

important and thus is a promising candidate for a single point mutation to reduce the flexibility of Pro68.

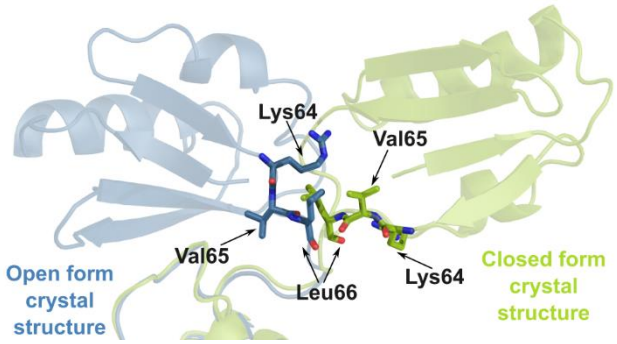


Figure 4.12. Zoom on the regulatory domain region of the X-ray crystal structures of the monomeric unit in its open (blue) and closed form (green). Lys64, Val65 and Leu66 are identified with sticks.

4.3.5 Key interactions in the tetrameric protein stabilise the regulatory domains in a closed state

If these three residues are essential for the flexibility of the linker region and by extension of the regulatory domains, a number of key individual residues might counterbalance them by forming strong non-covalent interactions between the regulatory domains and the rest of the protein in order to maintain the regulatory domains in a closed state. The analysis of the occurrence of hydrogen bonds between a regulatory domain and its adjacent counterpart, allosteric Tyr or the rest of the protein is reported in Table 4.1. Only the hydrogen bonds presenting an occupancy of more than 50%, in the trajectories for which the first 35 ns have been removed, are discussed hereafter.

Two adjacent regulatory domains in their closed conformation only form a hydrogen bond between His29 and Ser31 and a salt bridge between Glu35 and Arg46. This is somewhat

surprising given the stability of the closed form. The presence of Tyr enables the two adjacent regulatory domains to come together via the formation of various main chain hydrogen bonds with the main chain functionalities of Gln34, Glu35, Arg36 and Gly43, but also via a salt bridge between Asp45 and the ammonium ion of Tyr as well a hydrogen bond between the hydroxyl of Tyr and the hydroxyl of Ser31. Ser31 has been shown to be essential for the sensitivity to Tyr whereas His29 is responsible for the selectivity of Tyr versus Phe [257].

Table 4.1. Hydrogen bond occupancy between regulatory domains, between regulatory domains and allosteric Tyr and between regulatory domains and the rest of the protein. Only hydrogen bonds presenting an occupancy of more than 50 % over the course of an MD trajectory are reported. Residues that form salt bridges appear in red. The number of water molecules bridging two regulatory domains (more specifically involving residues 1 to 3 and 29 to 46 of the ACT domains which are either buried or responsible for the binding of Tyr) is averaged over the course of each simulation.

Donor	Acceptor	Closed_run1	Closed_run2	Closed_noTyr
Inter-regulatory domains hydrogen bonds				
ARG46-Side	GLU35-Side	84%	84%	86%
SER31-Side	HSE29-Side		53%	
Tyr - regulatory domains hydrogen bonds				
TYR-Main	ASP45-Side	88%	96%	
TYR-Side	SER31-Side	62%	86%	
TYR-Main	GLY43-Main	65%	86%	
GLN34-Main	TYR-Main	96%	98%	
GLU35-Main	TYR-Main	94%	97%	
ARG36-Main	TYR-Main	60%	74%	
GLY43-Main	TYR-Main	76%	95%	
TYR-Main	ILE41-Main		54%	
Regulatory domains - rest of protein hydrogen bonds				
ARG64-Side	GLU215-Side	91%	98%	81%
ARG247-Side	GLU54-Side	97%	97%	70%
SER55-Side	ALA306-Main	59%	77%	88%
TYR47-Side	GLU222-Side		74%	
ARG277-Side	ASP57-Side		71%	78%
LYS16-Side	GLU304-Side	63%	64%	79%
LYS311-Side	ASP57-Side	51%		
ARG36-Side	GLU75-Side	96%	95%	98%
LYS70-Side	GLU35-Side	78%		
Average number of water bridges linking two regulatory domains via hydrogen bonds				
Mean ± standard deviation		2.1 ± 1.0	2.6 ± 1.1	2.7 ± 1.2

Although the presence of Tyr is responsible for the formation of key interactions that indirectly link the two regulatory domains, this does not explain the stability of the closed form in the absence of Tyr. In addition, the absence of Tyr does not favour the presence of water molecules within the allosteric binding sites that could serve as a bridge between two regulatory domains as the average number of such kind of water bridges is the same for Closed_run2 and Closed_noTyr (Table 4.1). Moreover, each regulatory domain interacts with the rest of the protein via a number of hydrogen bonds and salt bridges but there are no major differences between the closed forms in the presence and absence of Tyr. Therefore, it seems reasonable to conclude that the predominant force stabilising the closed regulatory domains is the hydrophobic interactions between the two beta sheets of the regulatory domains, which are mainly composed of hydrophobic residues.

4.3.6 Differences in the flexibility of key catalytic loops are observed between the open and closed form simulations

To start addressing the impact of the domain movements on catalysis, the flexibility of individual residues was evaluated based on their Cα RMSF calculated on the monomeric unit of *T. maritima* DAH7PS by considering each chain composing the tetramer as being independent and combining them as individual trajectories and aligning all frames on the TIM barrel to remove undesired translational and rotational motions. The first 35 ns of each tetramer simulation were discarded as previously to allow for ample relaxation of the systems.

Unsurprisingly, the RMSF of residues 1 to around residue 63 is much more important for the open form simulations, accounting for the large amplitude motions observed for the ACT domain when simulated in their open state (Figure 4.13). Residues 64 to 67 are also more

flexible in the open trajectories, suggesting that the first residues of the linker region might play a more important role in the flexibility of the regulatory domains than the other linker region residues, confirming the findings made above regarding the importance of Arg64, Val65 and Leu66.

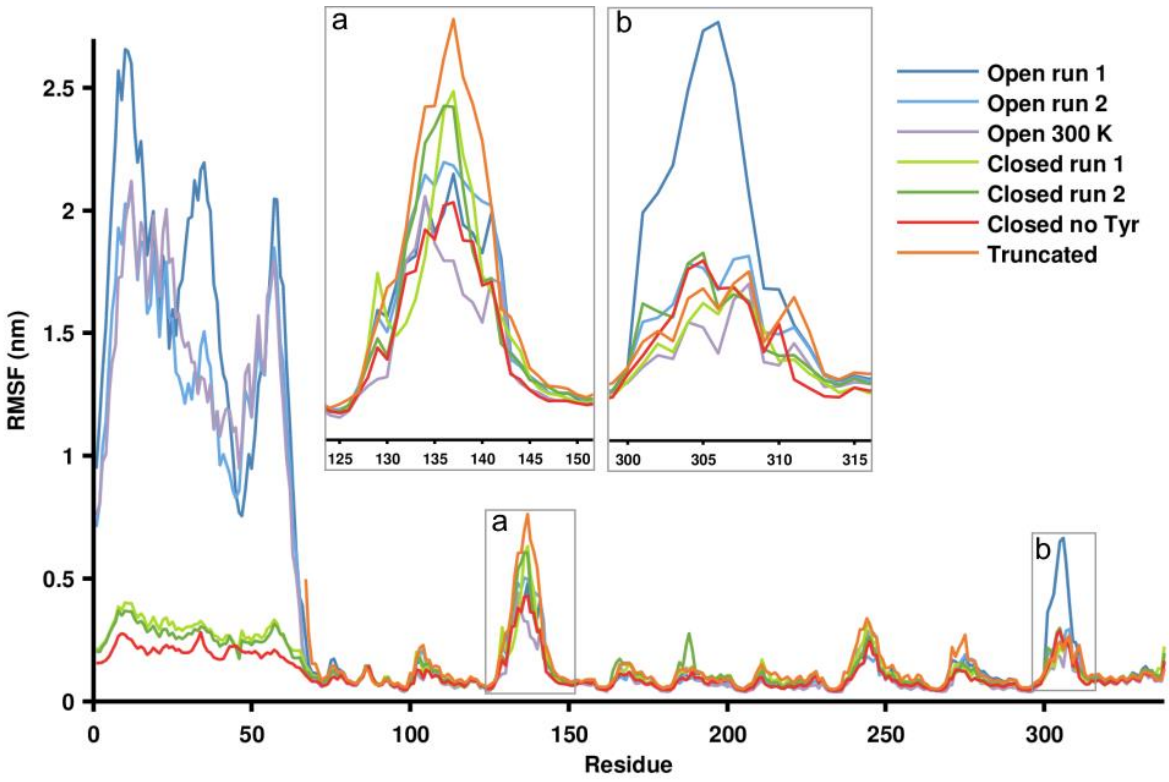


Figure 4.13. Cα RMSF of the MD trajectories of *T. maritima* DAH7PS after removal of the first 35 ns of each simulation. Open_run1 is represented in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green, Closed_noTyr in red and Truncated shown in orange.

Apart from the regulatory domains, the most flexible region corresponds to catalytic loop $\beta_2\alpha_2$ which is more flexible in the closed form simulations in the presence of Tyr when compared with the open form simulations, and almost reaches the degree of flexibility observed for the truncated mutant for which the movements of this loop are unrestrained. The position of this loop in the open form crystal structure is similar to its position in the closed form crystal structure. Careful examination of the closed form MD trajectories reveals that the $\beta_2\alpha_2$ loop

tends to rapidly leave the conformation adopted in the closed form crystal structure and move away, the longer the simulation, the further away the loop is (Figure 4.14 A). The same loop in the open simulations presents lower amplitudes, more random motions (Figure 4.14 B) and does not deviate as much from the open crystal structure.

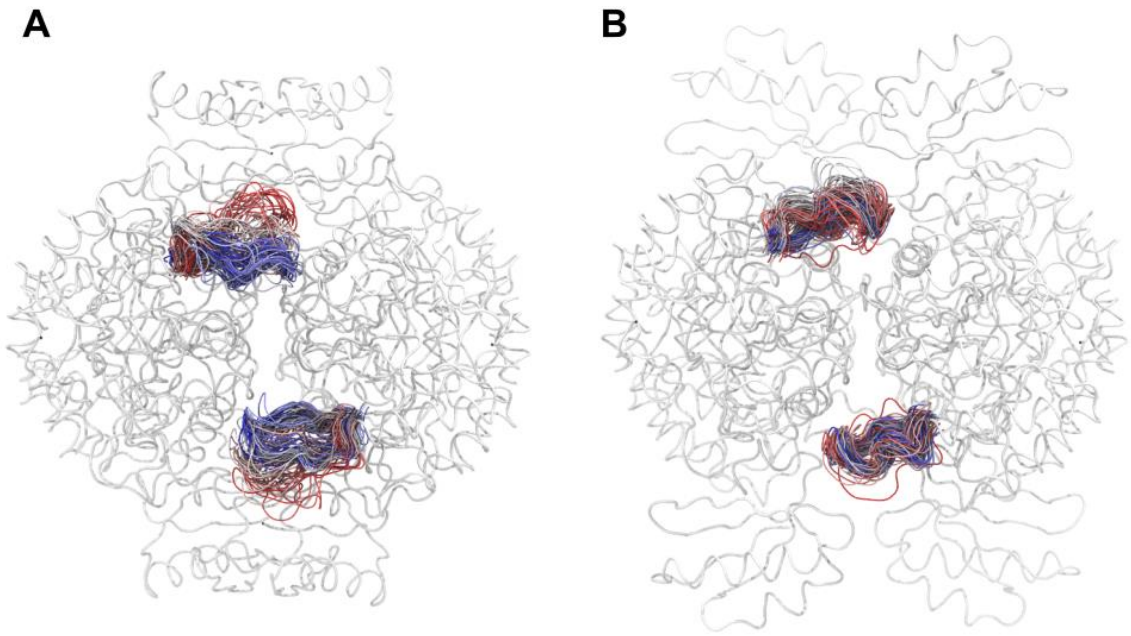


Figure 4.14. Snapshot of the $\beta_2\alpha_2$ loop motions taken at regular intervals over the course of the MD simulations for Closed_run2 (A) and Open_run2 (B). The loops are coloured with a gradient function of the increasing simulation time from red to white to blue.

The movement of the $\beta_2\alpha_2$ loop is confirmed by plotting the deviation of Pro136 Cα from the crystal structure as a function of time (Figure 4.15). Pro136 corresponds to the turn in loop $\beta_2\alpha_2$ and presents the greatest deviation. Globally the deviation of the loop in the closed form trajectories increases steadily before starting to plateau after about 150 ns, accounting for the gradually increased distance of the loop from its original position. For the open form trajectories, the distances tend to first increase and then decrease leading to lower deviations from the X-ray crystal structure than the closed form simulations, in agreement with more random motions of this loop.

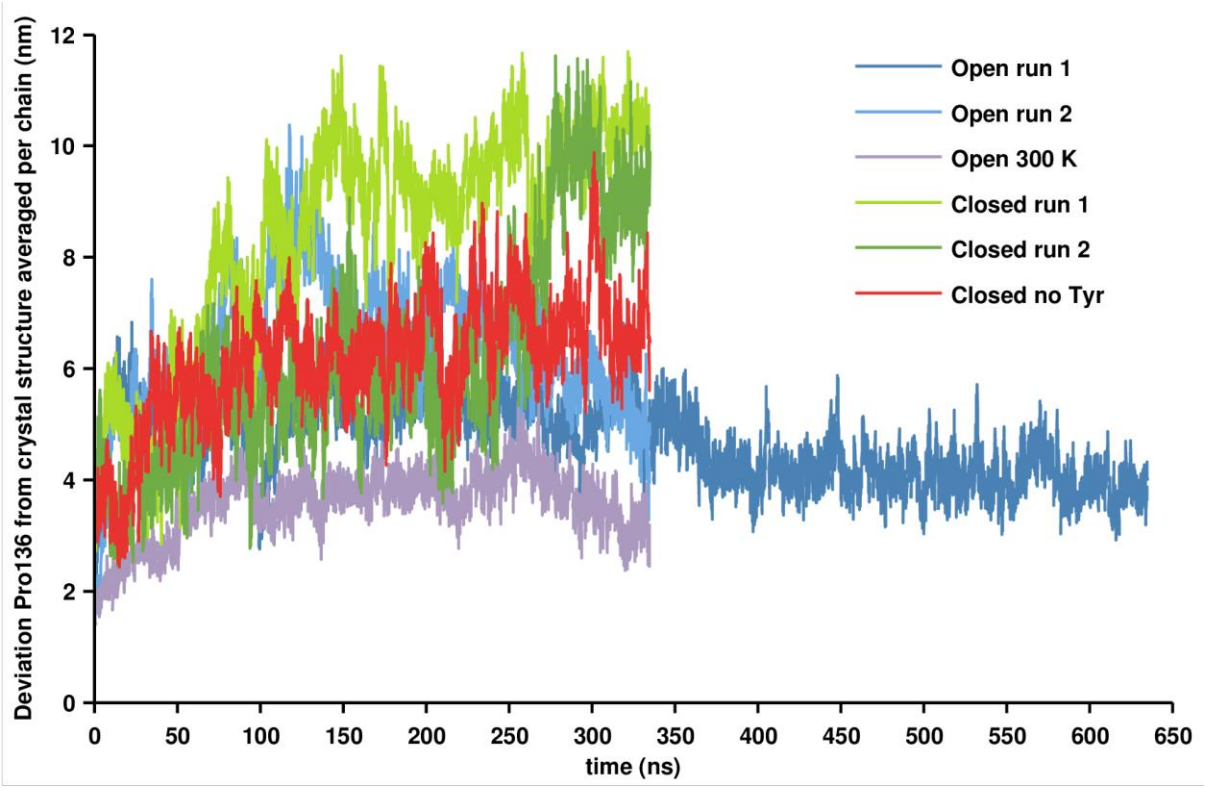


Figure 4.15. Deviation of Pro136 averaged by chains from its position in the crystal structure. Open_run1 is represented in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

In addition to the $\beta_2\alpha_2$ loop, loops $\beta_3\alpha_3$ (residues 165-170), $\beta_4\alpha_4$ (residues 184-190) and $\beta_5\alpha_5$ (residues 208-213) also seem to be more flexible in the closed state in the presence of Tyr or in the truncated mutant than in the open state or closed state in the absence of Tyr, but to a far lesser extent than the $\beta_2\alpha_2$ loop. These loops do not appear to be directly involved in catalysis, with the exception of Arg186, which binds a hydroxyl group of E4P (PDB 1RZM). Instead, they are at the interface with an adjacent monomeric unit and interact with other loops within the same monomeric unit or the adjacent one.

Catalytic loops $\beta_8\alpha_8$ and $\beta_7\alpha_7$ also exhibit differences between open and closed states. Overall in both cases, the flexibility of the loops seems to be higher for the open form simulations than for the closed ones, with a very high flexibility of loop $\beta_8\alpha_8$ observed for

Open_run1. This would indicate that these two loops have less freedom to move when the regulatory domains are closed and that this reduction of flexibility might be part of the inhibition mechanism. This is partly corroborated by the hydrogen bonds identified between this loop and the closed regulatory domains, in particular, Ala306 and Ser55, Lys 311 and Asp 57 as well as the salt bridge between Glu304 and Lys16 (Table 4.1).

Finally, loop $\beta_6\alpha_6$ which is involved in the binding of E4P and PEP presents a moderate increase in flexibility when compared with most of the other residues belonging to the main barrel, but no clear differences between open and closed forms can be identified.

4.3.7 The regulatory domains occlude the entrance to the active site by entering in contact with key catalytic loops

Analyses of the SASA, *R*g and RMSF as well as visualisation of the trajectories have revealed the regulatory domains hinder the entrance to the active site over the course of the open form simulations and the motion of $\beta_2\alpha_2$ loop in the closed form simulations lead to a more accessible active site. In order to quantify the occlusion of the active site by the regulatory domains both in the open and closed form trajectories, the number of contacts between the ACT domains and loops $\beta_2\alpha_2$, $\beta_6\alpha_6$, $\beta_7\alpha_7$, $\beta_8\alpha_8$ that form the contour of the active site were counted. Once a regulatory domain is in contact with one or more of these loops, the active site becomes occluded and the substrates have more difficulties to access it and, therefore, the activity of the enzyme can be affected. A loop is deemed to have one contact with a regulatory domain when any Cα of the regulatory domain falls below 10 Å of one alpha carbon of the loop, if more than one alpha carbon of the loop is affected, then the number of contacts increase accordingly. Therefore, the higher the number of contacts, the higher the number of residues

constituting the loop are affected. The distance of 10 Å was chosen as it gives the best selectivity and specificity for each loop of both the open and closed form simulations.

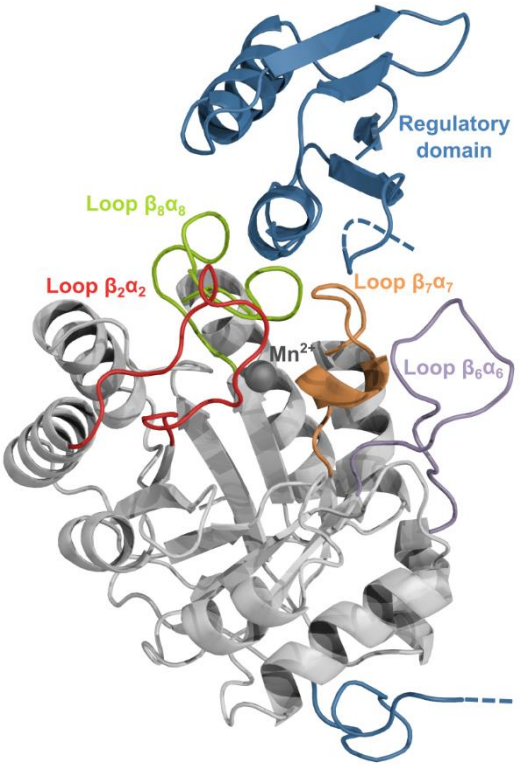


Figure 4.16. Structure of the catalytic barrel of one chain interacting with the regulatory domain of another chain in the closed form crystal structure. The active site of the enzyme is essentially formed by loops $\beta_2\alpha_2$ (red), $\beta_6\alpha_6$ (purple), $\beta_7\alpha_7$ (orange), $\beta_8\alpha_8$ (green). Regulatory domain is coloured in blue and the manganese ion in the active site represented with a grey sphere.

The average number of contacts per chain between each loop and the regulatory domains are plotted in Figure 4.17. For the closed form trajectories, the number of contacts between the regulatory domains and the $\beta_2\alpha_2$ loop rapidly decreases and can even fall below one, meaning that for some of the chains composing the tetramer, there is no contact at all between this loop and the regulatory domains. This is in line with the motions of this loop observed for the closed form simulations with this loop moving away from its original position. In contrast, the number of contact for this loop increases significantly over the course of the open form simulations at

363 K (the increase being less important at 300 K) and quickly exceeds the number of contacts in the closed form simulations, revealing that the motions of the regulatory domains lead to extensive contacts with the $\beta_2\alpha_2$ loop.

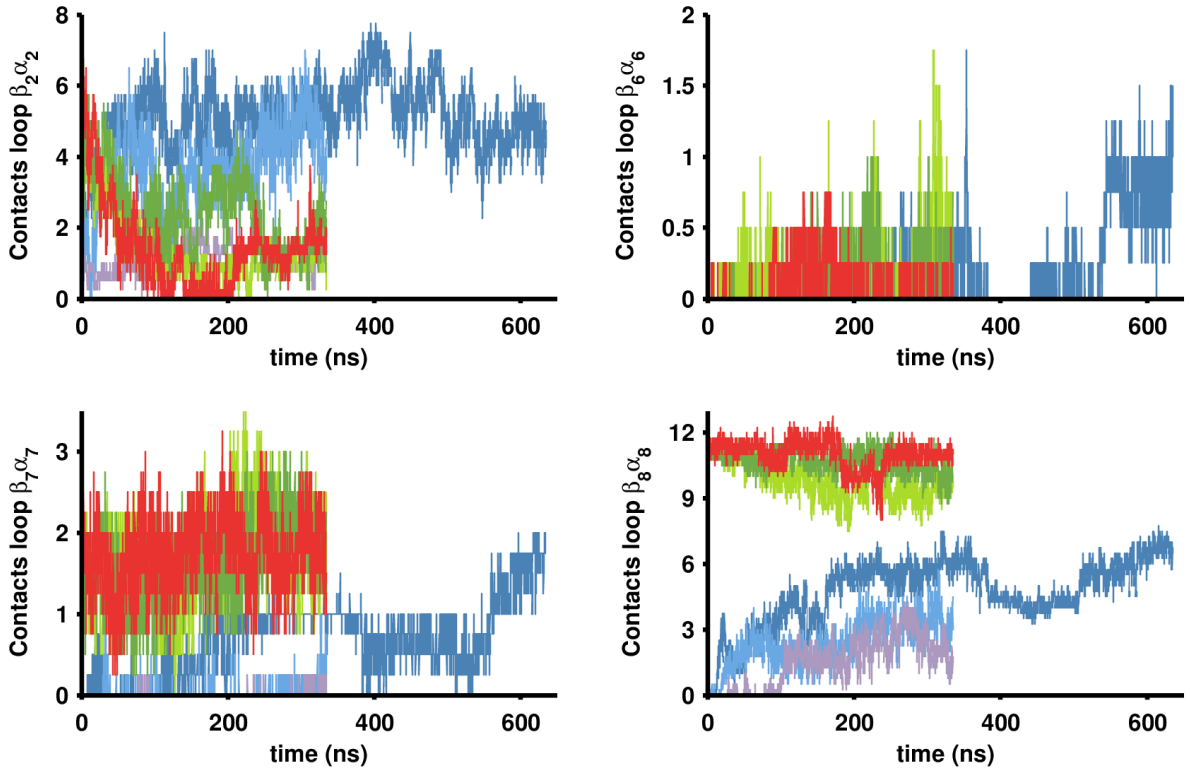


Figure 4.17. Number of contacts between residues of loops $\beta_2\alpha_2$, $\beta_6\alpha_6$, $\beta_7\alpha_7$, $\beta_8\alpha_8$ and the regulatory domain. The number of contacts has been averaged per chain. Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

The regulatory domains present few contacts with loop $\beta_6\alpha_6$ and loop $\beta_7\alpha_7$, which is unsurprising as these loops are the most distant ones from the regulatory domains due to the internal position of $\beta_6\alpha_6$ and the shortness of $\beta_7\alpha_7$. The contacts occur essentially over the course of the closed form simulations but also toward the end of the Open_run1 simulation. The number of contacts between loop $\beta_8\alpha_8$ and the regulatory domains remains very high during the closed form simulations accounting for the hydrogen bonds and salt bridges identified between this loop and the regulatory domains. Although initially not in contact with

this loop, the regulatory domains in their open conformation soon close the distance and a large number of contacts are observed as a result of the regulatory domain movements.

4.3.8 Displacement of the $\beta_2\alpha_2$ loop during the closed form simulations open a back door to the active site

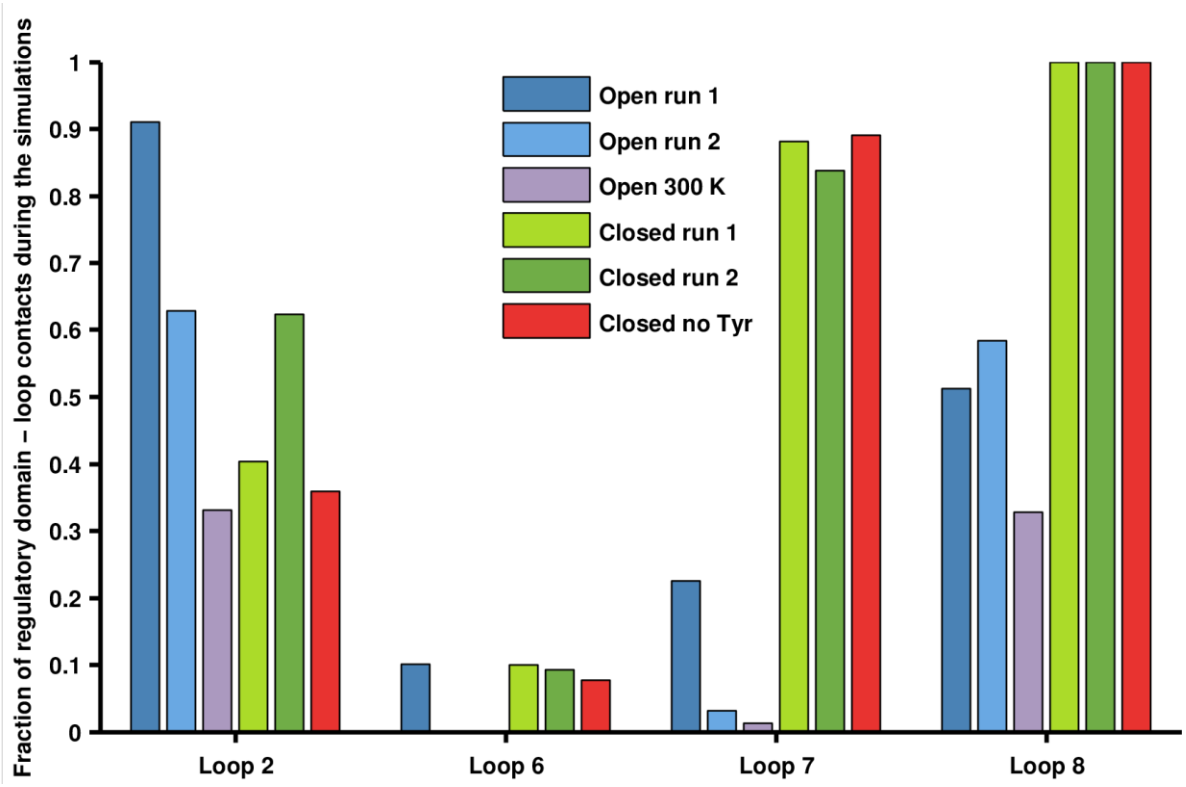


Figure 4.18. Fraction of the time for which at least one contact is observed between regulatory domains and a given loop. Open_run1 is shown in dark blue, Open_run2 in light blue, Open_300K in purple, Closed_run1 in light green, Closed_run2 in dark green and Closed_noTyr in red.

Looking at the fraction of the time for which at least one contact is observed between a regulatory domain and a given loop reveals that during the closed form simulations, the regulatory domains remain always in contact with loop $\beta_8\alpha_8$ and almost always in contact with

$\beta_7\alpha_7$ as revealed in Figure 4.18, which shows the fraction of the simulation time during which at least one regulatory domain – loop contact exists. However, contacts with loop $\beta_2\alpha_2$ are scarcer and almost non-existent in the case of loop $\beta_6\alpha_6$. The regulatory domains are in the majority of the cases at the same time in contact with loop $\beta_8\alpha_8$ and loop $\beta_7\alpha_7$ or with loop $\beta_8\alpha_8$, loop $\beta_7\alpha_7$ and loop $\beta_2\alpha_2$ (Table 4.2). Taken together, these results suggest that if blocking of the $\beta_8\alpha_8$ loop and, to a lower extent, the $\beta_7\alpha_7$ loop by the regulatory domains is a prerequisite to achieving inhibition of the enzyme, blocking of $\beta_2\alpha_2$ might be less critical and loop $\beta_6\alpha_6$ does not need to be blocked.

Table 4.2. Type of regulatory domains – loops contact existing at the same time during the MD simulations.

Type of contacts	*Open run 1*	*Open run 2*	*Open 300 K*	*Closed run 1*	*Closed run 2*	*Closed no Tyr*
No contact	1%	18%	34%	0%	0%	0%
Contact loop 2 only	48%	24%	33%	0%	0%	0%
Contact loop 8 only	3%	16%	32%	6%	9%	7%
Contact loops 2 & 8	26%	39%	0%	6%	6%	4%
Contact loops 7 & 8	2%	3%	1%	48%	23%	52%
Contact loops 6, 7 & 8	2%	0%	0%	5%	5%	6%
Contact loops 2, 7 & 8	10%	0%	0%	31%	53%	30%
Contact all loops	8%	0%	0%	4%	3%	2%

If the blocking of the $\beta_2\alpha_2$ loop is not deemed necessary to achieve enzyme inhibition, what effect has the displacement of this loop on the enzyme activity? Surface representation of the closed form simulations reveals that this loop movement opens a direct access to the active site (Figure 4.19). Not only this makes the active site solvent accessible, but the dimensions of the entrance can also accommodate the substrates PEP and E4P.

In order to verify that E4P and PEP can enter the active site via this back door, the two substrates were docked at the vicinity of the active site and both PEP and E4P were found to bind the entrance to the active site formed by this back door as well as its direct internal walls

(Figure 4.20). Therefore, this suggests that both the size of the back door is sufficient to accommodate the substrates, but also that the polarity of this active site entrance is adapted for the substrates to find their way inside the cavity.

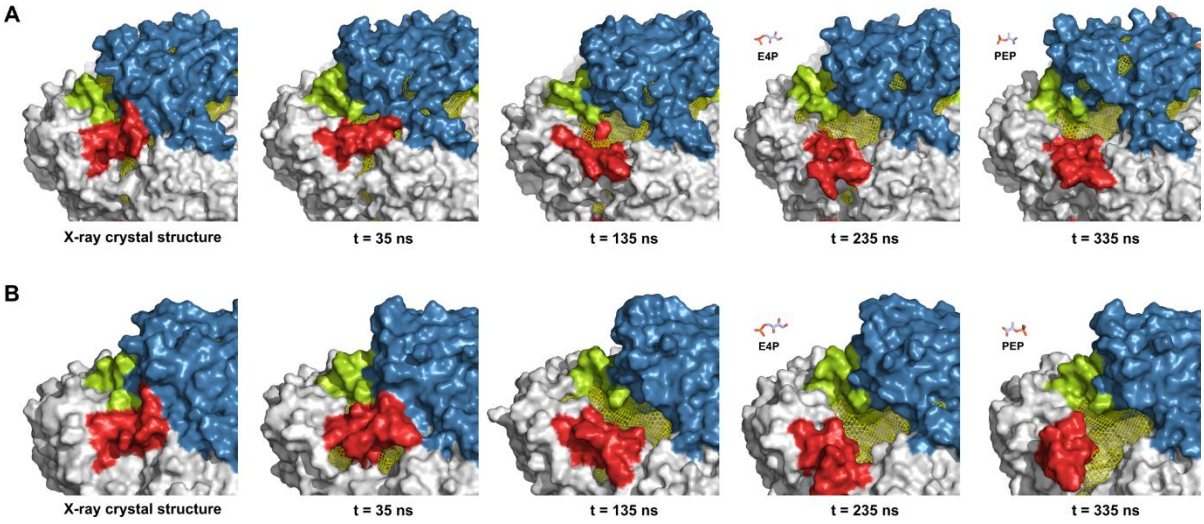


Figure 4.19. Molecular surface representation of the region corresponding to one active site and one regulatory domain during the MD simulations of Closed_run1 (A) and Closed_run2 (B). Regulatory domains are shown in blue, $\beta_2\alpha_2$ loop in red, $\beta_8\alpha_8$ loop in green and the rest of the catalytic barrel in light grey. The solvent exposed cavities are identified with a yellow mesh and were determined with KVFinder [266]. E4P and PEP are represented at scale with sticks to provide an indication of the size of the back door entrance.

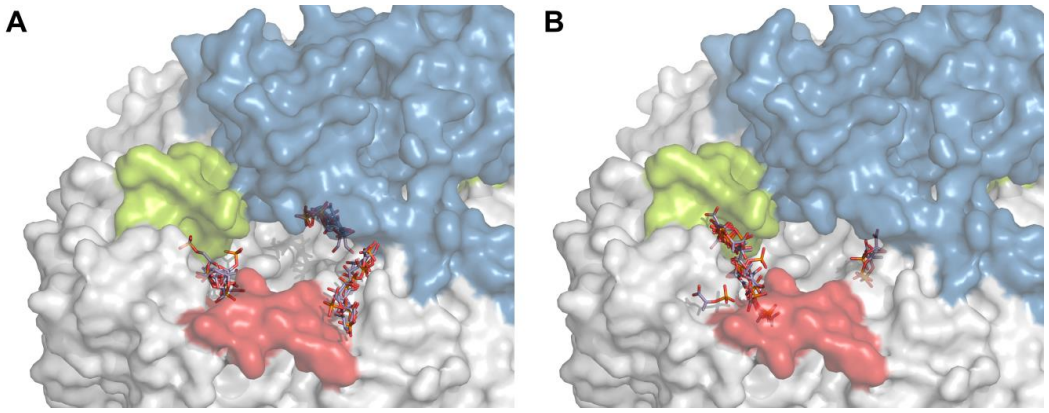


Figure 4.20. Docking of E4P (A) and PEP (B) to the back door entrance. The structure corresponds to the frame of Closed_run1 at 335 ns. Regulatory domains are shown in blue, $\beta_2\alpha_2$ loop in red, $\beta_8\alpha_8$ loop in green and the rest of the catalytic barrel in light grey.

4.3.9 Movements of the regulatory domains observed in the open form simulations impede the entrance to the active site

The open form simulations reveal that the $\beta_2\alpha_2$ loop is extensively in contact with the regulatory domains (Figure 4.18) explaining why this loop is significantly less flexible in the open form simulations than in the closed form simulations (Figure 4.13). Contrary to the contacts with the $\beta_2\alpha_2$ loop, which are present in the open form crystal structure, contacts with loop $\beta_8\alpha_8$ and to a far lesser extent with loop $\beta_7\alpha_7$ are created over the course of the simulations because of the large amplitude movements that are experienced by the regulatory domains due to the flexibility of the linker region (Figure 4.18). As shown in Table 4.2, during the open form simulations the regulatory domains can either have no contact at all with the loops, be in contact with only loop $\beta_8\alpha_8$, with only loop $\beta_2\alpha_2$ or with both at the same time. Interestingly, in the case of Open_run1, in 10% of the cases the regulatory domains are in contacts with loop $\beta_8\alpha_8$, loop $\beta_7\alpha_7$ and loop $\beta_2\alpha_2$ and in 8 % of the cases with all four loops. Consequently, in 18 % of the cases, the active site is completely occluded by the regulatory domains. This is illustrated in Figure 4.21, where the regulatory domains movement significantly compromises the open access to the top left active site after 35 ns. The active site remains, nonetheless, partly accessible but at the price of the substrates having to wander through the tortuous, constantly moving tunnel formed by the ACT domain interacting with the catalytic barrels and its adjacent counterpart. This is likely to affect both binding affinity of the substrates and release of the products.

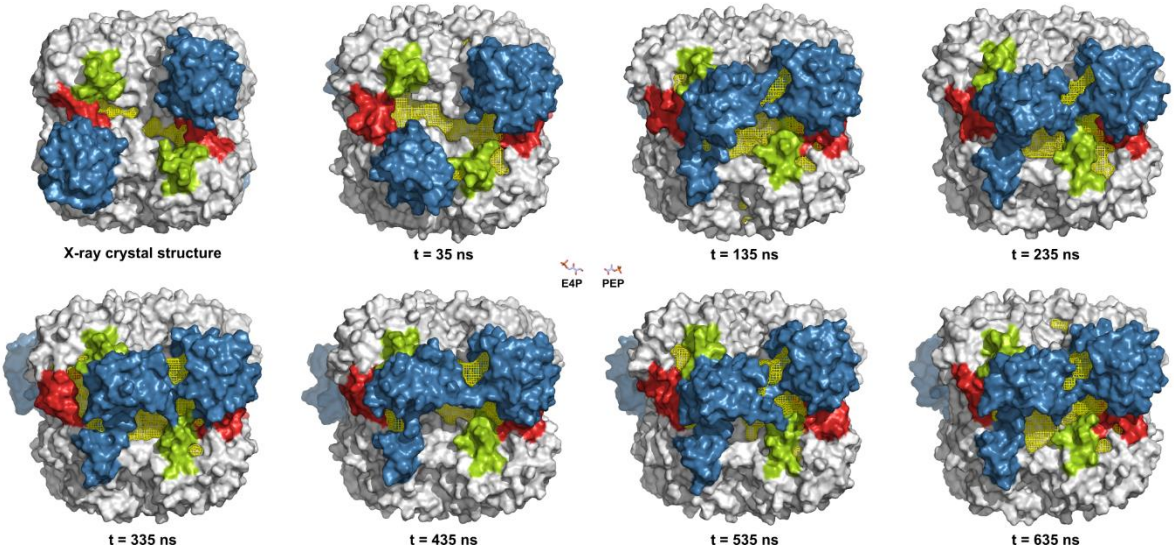


Figure 4.21. Molecular surface representation of the region corresponding to two active sites and two regulatory domains during the MD simulation of Open_run1. Regulatory domains are shown in blue, $\beta_2\alpha_2$ loop in red, $\beta_8\alpha_8$ loop in green and the rest of the catalytic barrel in light grey. The solvent exposed cavities are identified with a yellow mesh and were determined with KVFinder [266]. E4P and PEP are represented at scale with sticks.

4.4 Discussion

Using MD simulations, the dynamics in solution of tetrameric *T. maritima* DAH7PS, both in its open and closed forms, was studied in order to provide new insights into the allosteric regulation mechanism of this enzyme and of shed some light on some of the surprising experimental results observed.

The regulatory domains of the open form appear to be extremely mobile and do not retain the broadly open position captured in the crystal structure for more than a few nanoseconds before exploring the surroundings. This leads to the formation of transient interactions with the adjacent regulatory domain or with the catalytic barrels leading to a reduction in the solvent

accessibility of the main barrel. It is important to note that the region of the catalytic barrels that can be buried by the displacement of the regulatory domains corresponds, for the largest part, to the catalytic site of the adjacent chain. Therefore, any increased portion of the barrels buried by the regulatory domain diminishes the solvent accessibility of the active site of the enzyme. This can explains why the turnover number, k_{cat}, of the reaction is higher and K_m^{E4P} is lower for the truncated mutant than for the wild-type (WT) in the absence of Tyr (Table 4.3)[180]: the constant large amplitude motions of the regulatory domains impede the binding of E4P as well as the product release, thereby affecting k_{cat} and K_m^{E4P} compared with the truncated mutant. It is however important to note that the K_m of a single substrate presented here corresponds to an apparent K_m value measured experimentally by maintaining the concentration of the other substrate constant. Therefore, changes in the K_m value of a substrate might affect the K_m value of the other substrate. Although indicative of the affinity of a substrate for its binding site, the K_m are *not* a dissociation constants and the actual dissociation constants are unknown.

Table 4.3. Kinetic parameters for *T. maritima* DAH7PS. Kinetic parameters from [180]

	K_m^{PEP} (μM)	K_m^{E4P} (μM)	k_{cat} (s^{-1})
Wild-type in the absence of Tyr	4.85 ± 0.04	13 ± 1	11.7 ± 0.2
Wild-type in the presence of Tyr	21 ± 1	118 ± 10	2.59 ± 0.07
Truncated mutant	38 ± 3	6.6 ± 0.5	18.2 ± 0.3

The higher K_m^{PEP} for the truncated mutant is however somewhat surprising, but can possibly be explained by the results of the simulations. Firstly, the fact that E4P is bigger than PEP makes it more easily affected by the blocking of the active site by the regulatory domains as it can less easily and more slowly find its way through a tunnel leading to the active site as those formed during the regulatory domain motions (Figure 4.21).

In addition, in the truncated mutant, the motions of the loops, in particular, the $\beta_2\alpha_2$ loop, are completely unrestrained and it has been shown that an increase in flexibility of the catalytic loops can explain the reduction of activity observed upon binding of an allosteric ligand in type II [165] and type Iα DAH7PS [247]. Therefore, one possible explanation might be that the flexibility of the loops in the truncated mutant render the binding of PEP entropically unfavourable, leading to a slower *on*-rate and thus a higher K_m^{PEP} than for the WT (although as mentioned above the K_m may not represent the actual K_d and changes in K_d for E4P may also impact K_m^{PEP}) for which the motions of the loops are restrained by the regulatory domains. Moreover, due to the accessibility of the active site in the truncated mutant, the *off*-rate of PEP binding may be higher than in the WT enzyme as the domain motions are expected to slow the release of the substrate and products from the active site. Binding of PEP might then form stabilising interactions that reduce the flexibility of the loops, making the binding of E4P much easier. In contrast, the flexibility of the loop is lower in the WT even in the absence of PEP, so the transient blocking of the entrance to the active site is primarily responsible for a reduced binding affinity. In the case of E4P, the *off*-rate is less affected because PEP is likely to have already bound the active site, making the reaction between the substrate directly possible whereas in the case of PEP, the substrate has to wait for the E4P to bind.

Although the closing of individual ACT domain is relatively facile, the closing of all four domains is expected to be a much rarer event and, therefore, to occur on a longer timescale. What is, however, the role of Tyr in the closing of the regulatory domains? Is the enzyme able to attain the fully closed form in the absence of Tyr and binding of Tyr stabilise this form (conformational selection mechanism)? This is unlikely, as simulations of the closed form in the absence of Tyr have proven to be very stable. Indeed, the closed form is stabilised by both strong hydrophobic interactions at the interface of the two β-sheets composing the ACT

domain, and the hydrogen bonds and salt bridges existing between the two ACT domains and between the ACT domains and the catalytic barrels. Tyr seems to form a number of interactions between the two ACT domains but these interactions do not seem to be required for the enzyme to remain in position and, therefore, we can conclude that the stabilising role of Tyr is low relative to the strong forces that maintain the ACT domain in a closed conformation. This is also confirmed by the fact that release of a Tyr does not trigger the opening of the regulatory domains, as observed for Closed_run1 (Movie 4.2) and thus the absence of Tyr does not seem to be primarily responsible for the opening of the ACT domains. Then does Tyr trigger the closing (induced fit mechanism)? Not exactly as the individual domains can explore a large number of conformations, including the partially closed conformation easily. The most likely explanation is that the regulatory domains can sample a large number of conformations and interact with each other but cannot reach easily the fully closed form because of high energy barriers. The role of Tyr might be to promote certain interactions between adjacent regulatory domains that are not often sampled, thereby leading to a more facile closing of all the regulatory domains. Tyr can for example, assists in the correct positioning of two adjacent ACT domains by engaging non-covalent interactions with one and then serving as a bridge to recruit the second ACT domain in a position intermediate to the closed form. This situation is illustrated in Figure 4.22. The fully closed conformation might be virtually inaccessible because of high energy barriers, Tyr might lower the energy and the barriers of some key intermediate states in the closing process, and thus facilitates the conformational transition to a fully closed form.

One question remains: what is the role of the displacement of the $\beta_2\alpha_2$ loop in the closed form? A possible scenario could be that the increased flexibility of the $\beta_2\alpha_2$ loop decreases the binding affinity for the substrates or decreases the catalytic activity. This is unlikely as the binding affinity would be at its lowest value by maintaining the loop closed, as in the crystal

structure, and the catalytic activity of the truncated mutant is high despite the flexibility of this loop. Moreover, if the $\beta_2\alpha_2$ loop appears to have a high RMSF, this is due to its overall displacement rather than the fluctuations of this loop, so the situation is different than a constantly moving loop as it is the case in type II, type Iα or in the truncated mutant. Indeed, once the $\beta_2\alpha_2$ loop is open, it remains in this position and does not fluctuate anymore.

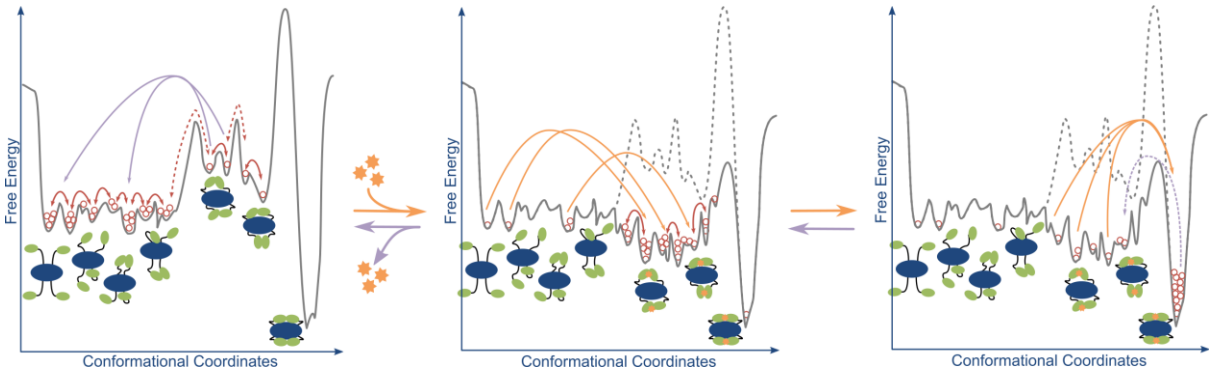


Figure 4.22. Schematic representation of the energy landscape for the closing and opening mechanism of *T. maritima* DAH7PS. Each bead represents the relative population of a given state and the red arrows represent the classical conformational changes the protein experience due to its dynamic. Dashed arrows indicate infrequent events. In the absence of allosteric ligand, the regulatory domains (green ovals) explore the swallow energy landscape and interact with the adjacent regulatory domains and the main barrel (blue oval). Most of these interactions are not real intermediate states of the closed state but can interconvert rapidly due to low energy barriers. A few conformations can serve as intermediates to give access to the fully closed state, but these conformations are not often sampled due to higher energy barriers, and the fully closed form remains essentially inaccessible. The addition of Tyr (orange stars) enables to stabilised these conformations by reducing their relative energy and the associated barriers. As a result, there is a population shift toward these states (orange arrows, middle panel) upon binding of Tyr and these intermediate states can be sampled more often and thus, the fully closed state is more readily accessible due to its lower energy barrier, permitting another population shift from the intermediate states to the fully closed state (orange arrows, right panel). Unbinding of Tyr might not occur directly form the fully closed state. Instead, the opening of the $\beta_2\alpha_2$ loop might be able to destabilise the fully closed form, leading to the sampling of intermediate conformations (purple arrow, right panel). Then unbinding of Tyr can occur from these intermediate conformations, and upon unbinding, these conformation becomes of high energy and the system return to the set of conformation initially sampled (purple arrows, left panel).

Therefore, the displacement of this loop is more likely to open a back door, which may serve two potential purposes: it might restore partial activity or permit the opening of the regulatory domains. It has been shown that this new entrance is large and is likely to be able to accommodate the substrates, enabling the enzyme to perform its reaction despite the regulatory domain "blockade". This would explain the remaining activity of the enzyme observed experimentally when fully inhibited and because E4P is bulkier than PEP, its entrance not as facile and thus K_m^{E4P} is higher than K_m^{PEP} (with the proviso that these K_m values are not dissociation constants as mentioned previously). However, for the enzyme to be able to perform this catalysis, the closing of the back door is required so catalytic residues belonging to the $\beta_2\alpha_2$ loop can serve their purpose. As the loop seems to be stable in this open state, the closing might be a limiting step to restore full catalytic activity. Moreover, after the reaction, the product is much bulkier than the two separate substrates and, therefore, release of the product might be slower than in the active open enzyme. This two factors might explain why the activity remains very low despite the substrates being able to enter the active site.

As observed during the simulations, the two regulatory domains in a closed state form very stable interactions, which are not easily broken even in the absence of allosteric ligand. Therefore, another possible role, which might happen in parallel, is that opening of a back door might disrupt the non-covalent interactions between the main barrels and the regulatory domains (for example via the interactions with water molecules that enter the region buried by the regulatory domains), destabilising the closed form and facilitating the opening. Such a mechanism is observed in the Closed_run1 trajectory (Movie 4.2), in which the displacement of the $\beta_2\alpha_2$ loop seems to lead to the cleavage of the interactions as well as the loss of contacts between the $\beta_8\alpha_8$ loop and its ACT domain partner. In turn, this enables a partial opening of the two regulatory domains, allowing one Tyr to escape its allosteric binding site. If the same

loss of contact between the adjacent regulatory domain and its interacting $\beta_8\alpha_8$ loop would have happened, it is likely that it would have resulted in the two regulatory domains to break apart, releasing the second Tyr. Therefore, it is likely that the opening of the regulatory domains is not, at least initially, triggered by the unbinding of Tyr. Instead, the opening of the $\beta_2\alpha_2$ loop can eventually lead to the destabilisation of the closed form, and the enzyme can readopt a conformation similar to the intermediate conformations mentioned previously that are involved in the closing mechanism (Figure 4.22), and it is from these conformations that Tyr can dissociate from the regulatory domains which in turn leads to the adoption of the sampling of open conformations.

In conclusion, the regulatory mechanism of *T. maritima* DAH7PS is not straightforward and presents several hidden complexities, such as the exact role played by the allosteric ligand in the opening mechanism of the regulatory domains as well as the role of the displacement of the $\beta_2\alpha_2$ loop in the residual activity of the enzyme in the presence of ligand as well as in the opening mechanism. Moreover, if the large, rather clumsy, domain motions seem like a gross and ineffective way to regulate activity, it is also an easy way to implement allostery in an unregulated protein by recruitment of an external domain via gene fusion. The regulation can then be rapidly fine-tuned via evolution of the residues forming interactions between the regulatory domains and the catalytic barrels, and of the residues composing the linker region in order to provide the correct level of flexibility.

Chapter 5 – The allosteric regulation of type Iβ DAH7PS from *Geobacillus sp.*

5.1 Introduction

From sequence analysis, the DAH7PS from *Geobacillus sp.*, a thermophilic soil bacterium, is predicted to have an N-terminal CM domain and be a member of the type Iβ group. The independence of the individual catalytic DAH7PS and CM has already been demonstrated for the *Bacillus subtilis* enzyme [210] and a structure of an uninhibited form enzyme of this type has been previously determined for the DAH7PS from *Listeria monocytogenes* [181]. This structure shows two homodimeric CM domains extending from either side of the catalytic barrel tetrameric plane. In this chapter, the structural changes associated with the allosteric response to prephenate are studied using X-ray crystallography and small angle X-ray scattering (SAXS) experiments and it is found that prephenate binding to the CM domain stabilises a more compact conformation of the protein, limiting the catalytic functionality of the DAH7PS domain active site. In addition, the on-going molecular dynamics simulations of *Geobacillus sp.* DAH7PS (*Gsp*DAH7PS) that should reveal additional information on the dynamics of this enzyme and on its allosteric regulation mechanism are introduced.

The first part of this chapter, dedicated to the analysis of the X-ray crystal structure, homology models and SAXS data are extracted from an article in preparation:

Nazmi AR, Othman M, Lang EJM, Bai Y, Allison TM, Panjikar S, Arcus VL, Parker EJ: **Tipping an allosteric hat: The catalytic regulatory domain caps the active site of 3-deoxy-D-*arabino* heptulosonate synthase from *Geobacillus sp.*,** *In preparation 2016*

Dr Ali Reza Nazmi and Mohamad Othman designed experiments, expressed and purified the enzyme and conducted its biochemical characterisation (not reproduced here with the exception of the kinetic parameters). Dr Ali Reza Nazmi crystallised the enzyme and refined the crystal structure. Dr Timothy M Allison, Dr Santosh Panjikar and Prof Vickery L Arcus assisted with X-ray crystallography. Yu Bai expressed and purified the enzyme, prepared the SAXS samples, and analysed the SAXS data and contributed to the SAXS section of the article. Prof Emily J Parker, designed experiments, collected the SAXS data and wrote the manuscript. Dr Ali Reza Nazmi and Dr Timothy M Allison contributed to the manuscript.

I contributed to the final stages of refinement of the crystal structure, designed the molecular modelling experiments, constructed the model of the inhibitor-free structure by homology with *Listeria monocytogenes* DAH7PS (*Lmo*DAH7PS) as well as modelled the missing residues of *Gsp*DAH7PS. I analysed the crystal structure, the models and participated in the interpretation of the SAXS data. I wrote the section of the article on the analysis of the crystal structure and of the models and contributed to the SAXS section. This work is described in this chapter.

5.2 Methods

5.2.1 Reconstruction of the missing residues and side chains of GspDAH7PS

Missing residues and side chains from the structure of *Gsp*DAH7PS obtained after refinement were reconstructed using MODELLER [270] version 9.13. Firstly, the tetrameric form was generated from the asymmetric dimeric unit in PyMOL [271] and the topologies and parameters for prephenate and Mn^{2+}, obtained from the SwissParam server [272] and the MM3 force field [273-275] respectively, were added to the MODELLER library. Secondly, 100 models of *Gsp*DAH7PS were generated using the "automodel" routine of MODELLER tuned to improve model accuracy. More specifically the variable target function method was set to slow and the maximum number of conjugate gradient iterations set to 300 while the level of molecular dynamics refinement was adjusted to slow and the complete optimization-refinement process was repeated twice. The `env.io.hetatm` instruction defined as true in order to include the metal ion and prephenate in the model. During the model generation, only the missing residues and residues with missing atoms were allowed to move, the chorismate mutase domains were restrained to keep their secondary structure intact and a symmetry restraint was applied on the alpha carbon of the symmetric chains of the tetramer. The best model was selected among the models presenting the lowest MODELLER objective function value and the lowest global DOPE score, following visual inspection and residue-by-residue DOPE score analysis. The C-terminal tail of this model was further refined using the "loopmodel" class of MODELLER [276] using the same protocol as describe above and the best of the 100 models generated was assessed using the same evaluation procedure.

5.2.2 Homology model construction of GspDAH7PS in its open form

Homology modelling of *Gsp*DAH7PS in its open active form was performed using the two crystallised structures of DAH7PS from *Listeria monocytogenes EGD-e* [Protein Data Bank (PDB) ID code: 3NVT and 3TFC] as templates. For each structure of *Lmo*DAH7PS, the tetrameric form was generated in PyMOL, and the topology and parameters of PEP obtained with SwissParam as well as those of Mn^{2+} were included into MODELLER libraries. The missing residues and side chains were then added with MODELLER using "automodel" to create 25 models using the same parameters and the same assessment procedure as described above. The best model for each of the structures were aligned and their sequences were aligned against the full sequence of *Gsp*DAH7PS using T-Coffee [224]. The homology model of *GspDAH7PS* in its open form was performed with MODELLER using the "automodel" routine as before with the `env.io.hetatm` instruction was set to true in order to include PEP and Mn^{2+}in the model. Secondary structure restraints were applied to the chorismate mutase domains and symmetry restraints were used for the symmetric chains of the tetramer. A total of 100 models were generated, the best model was selected using the previously reported method.

5.2.3 Analysis of the small-angle X-ray scattering data

Scattered intensity (I) was plotted versus s using Primus [277]. All samples were devoid of an increase in intensity at low s (indicative of aggregation). Guinier plots were linear for $sR_g < 1.3$. The data sets for structural analysis were recorded with 447 data points over the range $0.005 \leq s \leq 0.35$ Å^{-1}. Theoretical scattering curves were generated for the models of the open active and closed inactive forms of *Gsp*DAH7PS and compared with experimental scattering curves using OLIGOMER [277]. Theoretical scattering data of two distinct *Gsp*DAH7PS

forms were further mixed and fitted to the individual experimental scattering profile in presence or absence of prephenate with OLIGOMER to investigate the volume fraction of each form in the solution with and without ligand [277].

5.2.4 Additional computational methods

Solvent accessible surface area (SASA) for the open and closed models were calculated in VMD [265] using an in-house Tcl script and using a solvent probe radius of 1.4 Å. Sequence identity was determined using EMBOSS Matcher [278,279]. Sequence alignments were generated with T-Coffee [224,280] and rendered with ESPrit 3 [281]. Structural figures were generated with PyMOL.

5.3 Results

5.3.1 Prephenate binds to the chorismate mutase domain at a binding site located within a helical bundle

*Gsp*DAH7PS was crystallised in the presence of Mn^{2+} and chorismate and its structure was determined at a resolution of 2.75 Å in the space group P64. The asymmetric unit is composed of two adjacent chains, with each comprising a catalytic $(\beta/\alpha)_8$ barrel common to all DAH7PS enzymes (residues 106 to 362), a CM domain (residues 1 to 86) connected by a short linker region (residues 86 to 105), as shown in Figure 5.1. The two catalytic barrels from each chain in the asymmetric unit are highly similar (Cα RMSD = 0.11 Å) whereas the CM domains are slightly less similar (Cα RMSD = 0.23 Å), with one of the chains having more unresolved

residues than the other one in the linker region connecting the CM domain with the catalytic barrel. However, when taken as a whole, the two chains present a Cα RMSD of 3.93 Å. This large difference arises from the difference in placement of the two CM domains relative to the barrels (Figure 5.1A).

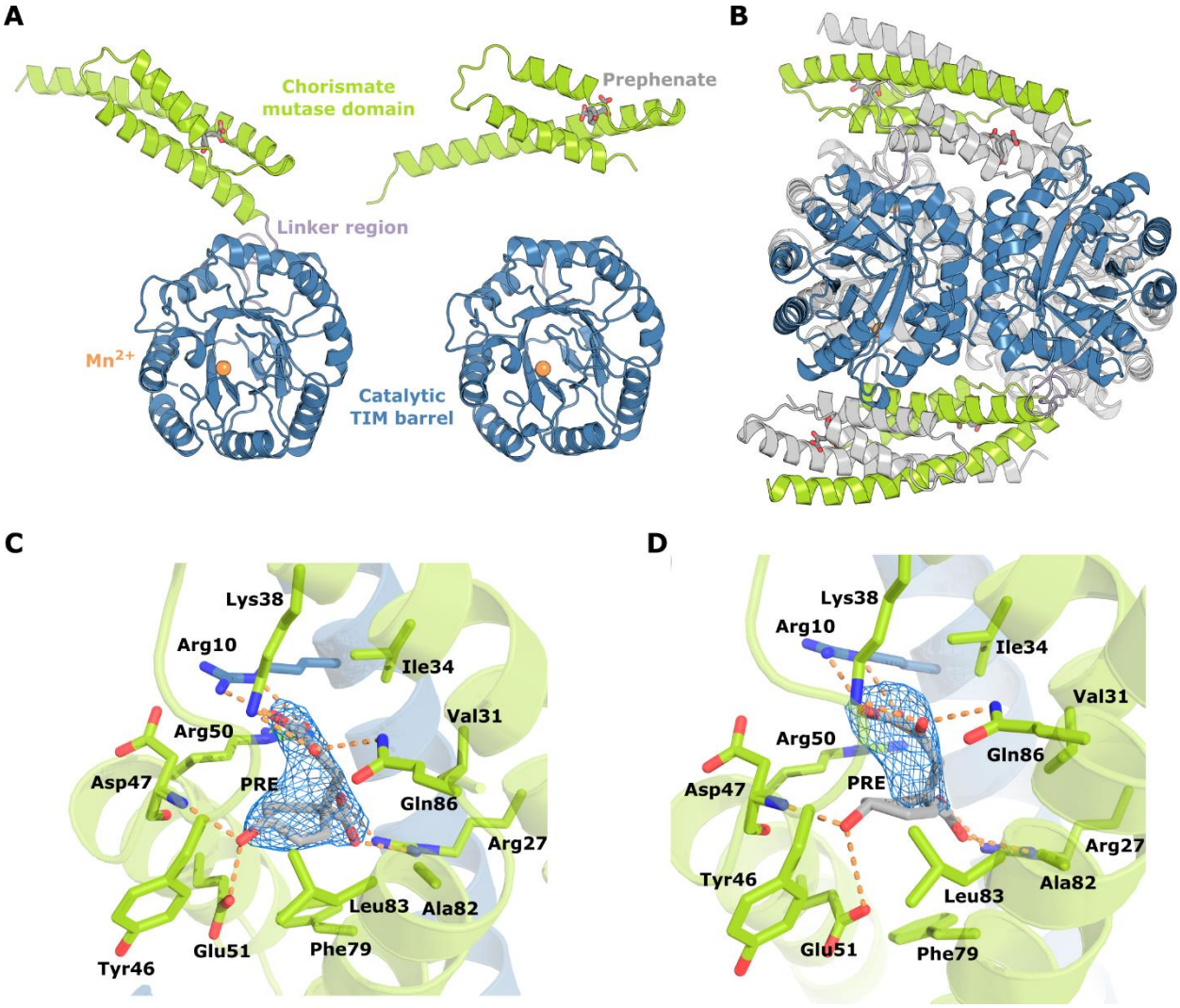


Figure 5.1. Crystal structure of *Gsp*DAH7PS. A. Chain A (left) and chain B (right) constituting the asymmetric unit. The catalytic TIM barrel is shown in blue, the CM domain in green and the linker region in purple. Allosteric inhibitor prephenate (PRE) is represented with grey sticks while the active site is identified with the metal ion, represented by an orange sphere. **B.** Tetrameric assembly with one asymmetric unit coloured as previously and the other one coloured in grey. **C.** and **D.** Binding sites of prephenate. Residues important for the binding of prephenate and the catalytic activity of the CM domain are shown with sticks. One CM domain is shown in green and the adjacent chain forming the catalytically active CM is shown in blue. Prephenate is represented with grey

sticks. The Fo-Fc omit maps for prephenate are represented with a blue mesh and were generated by deleting the prephenate molecules from the n-1 round of refinement the density is contoured at contoured at 3σ (0.1457e/A3).

A two-fold symmetry operation reveals the homotetrameric biological unit of *Gsp*DAH7PS. The DAH7PS barrel subunits are arranged in a similar fashion to the tetrameric assemblies observed for other structurally characterised type Iβ DAH7PS enzymes [177], with neighbouring protomers placed so that adjacent barrels face in opposite directions around the tetramer plane. A manganese ion is coordinated at the C-terminal end of each of barrels the by residues Cys126, His296, Glu322 and Asp333, depicting the locality of the active site. These metal binding residues are conserved in all structurally characterised DAH7PS enzymes. The helical CM domains of two diagonally opposed chains associate in a tightly interlaced dimer that resembles a type II or AroQ helical CM [238].

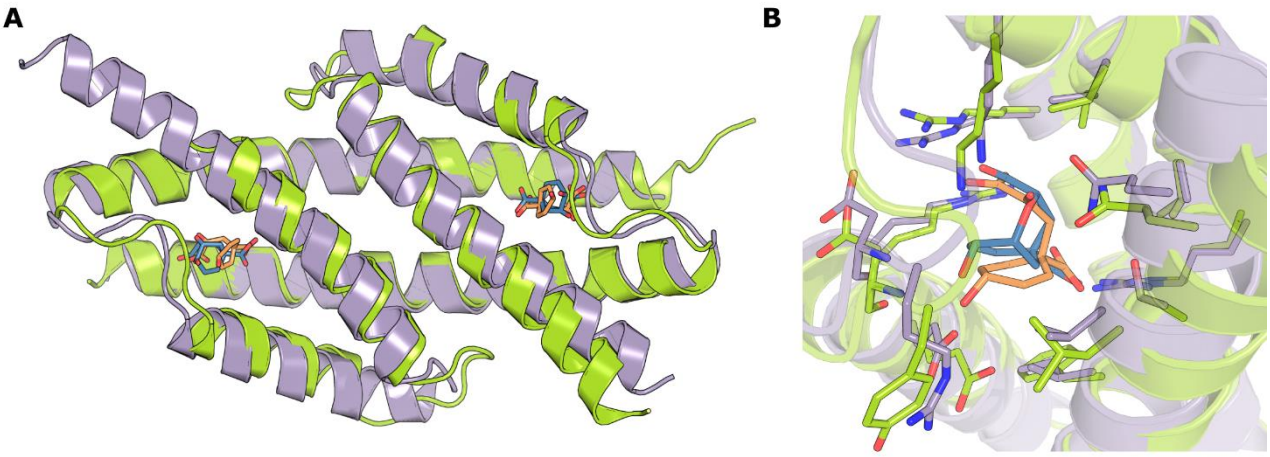


Figure 5.2. Comparison between the CM domains of *Gsp*DAH7PS and *Eco*CM. **A.** Structural alignment of the CM domains of *Gsp*DAH7PS (residues 1 to 94) and *Eco*CM. *Gsp*DAH7PS is shown in green with prephenate in orange sticks and *Eco*CM is shown in purple with the transition state intermediate in blue sticks **B.** Superimposition of the CM active site.

Despite low sequence identity (34.5%) the CM domains of *Gsp*DAH7PS align well with *E. coli* AroQ CM [238], (*Eco*CM, RMSD = 2.23 Å) (PDB code: 1ECM, Figure 5.2). The

*Gsp*DAH7PS CM helices show more curvature than is observed in the *Eco*CM structure, especially noticeable at the end of the first helix (Figure 5.2), which may be the result of the interactions between the CM domains and the DAH7PS main barrels. Moreover, the second and third helices of the *Gsp*CM domains are shorter than their *Eco*CM counterparts.

Electron density was observed in the area corresponding to the catalytic site of the CM domain. Due to the observed CM activity of *Gsp*DAH7PS, it is likely that the added chorismate co-crystallised with the protein was converted into prephenate and, therefore, the prephenate was modelled into the CM domain active site (Figure 5.1 C and D). Modelling of prephenate within the active site reveals that the residues involved in its binding correspond to a subset of the residues responsible for the catalytic activity of the CM domain (Figure 5.2 C). More specifically, the position of Arg10, Arg27, Lys38, Asp47 and Glu51 suggest they form hydrogen bonds and other electrostatic interactions with prephenate whereas Val31, Ile34, Met54, Phe79 and Leu83 create a hydrophobic shell surrounding the ligand, and the aliphatic side chain of Arg50 contributing to these hydrophobic interactions.

The well-characterised catalytic site of *Eco*CM [237-240] exhibits an excellent structural alignment with the active site of the CM domain of *Gsp*DAH7PS (Figure 5.2 A), with the catalytically important charged and polar residues shared by both proteins. Hydrophobic residues Ala32, Val35, Ile81 and Val85 of *Eco*CM, are not conserved in *Gsp*DAH7PS but instead replaced by alternative hydrophobic residues: Val31, Ile34, Phe79 and Leu83 respectively. In addition, Ser84, which had been postulated as being important in *Eco*CM [237] is not conserved and replaced by Ala in *Gsp*DAH7PS. These alterations may account for the relatively low CM activity of the *Gsp*DAH7PS^{WT} and the truncated protein.

5.3.2 Prephenate binding is associated with closer association between the catalytic and regulatory domains

The structure of an N-terminal CM domain regulated DAH7PS has previously been reported for *Listeria monocytogenes* (*Lmo*DAH7PS, PDB IDs: 3NVT and 3TFC) [181]. This structure, however, differs from that presently reported for *Gsp*DAH7PS in the sense that it is an active, uninhibited crystal structure of DAH7PS in the presence of substrate PEP. The position of the CM domains in the *Lmo*DAH7PS structure are found further from the catalytic DAH7PS barrels in an "open" position when compared with the "closed" position they adopt in the inhibited prephenate-bound *Gsp*DAH7PS structure. *Lmo*DAH7PS shares sequence identity of 77% with *Gsp*DAH7PS (Figure 5.3) and the structural alignment of tetrameric *Gsp*DAH7PS with *Lmo*DAH7PS yields RMSD values of 0.82 Å (3NVT) and 1.08 Å (3TFC).

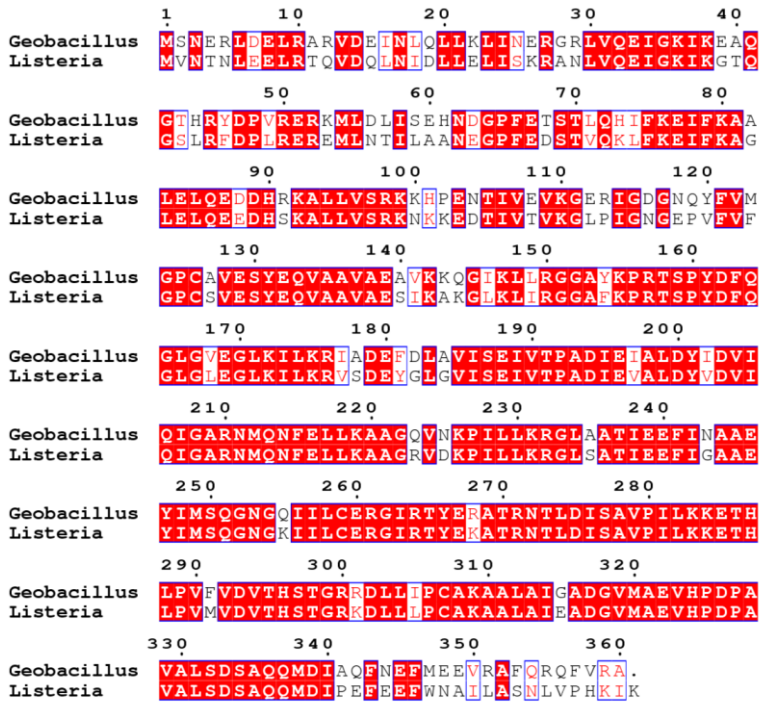


Figure 5.3. Sequence alignment of *Lmo*DAH7PS and *Gsp*DAH7PS.

In order to estimate the likely structural changes of *Gsp*DAH7PS on the binding of prephenate, a model of *Gsp*DAH7PS in its active open form based on the two known *Lmo*DAH7PS structures was created. The homology model was created using MODELLER [270] after having reconstructed the loops missing in the two *Lmo*DAH7PS structures. A total of one hundred models were generated and the best model was selected on the basis of the lowest MODELLER objective function value and the lowest global DOPE score, followed by analysis of the residue-by-residue DOPE score and visual inspection of the structure. Additionally, the residues and side chains for which electron density is not observed from the crystal structure of *Gsp*DAH7PS were added in order to generate a complete model of the inactive closed form of *Gsp*DAH7PS. The resulting models are presented in Figure 5.4.

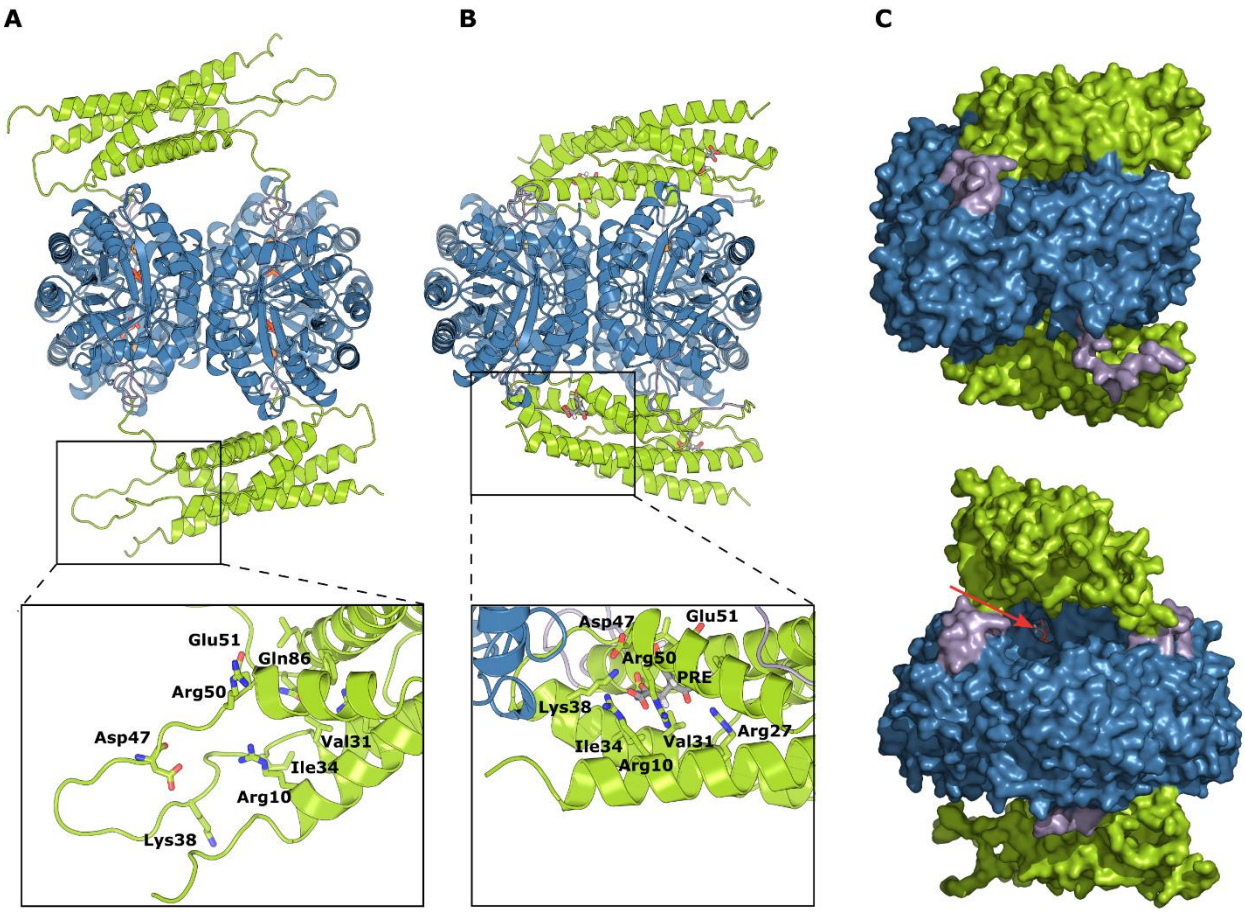


Figure 5.4. Open (A) and closed (B) models of *Gsp*DAH7PS corresponding respectively to the apo and allosterically inhibited states of the enzyme. The open model consists in a homology model based on the deposited structures of *Lmo*DAH7PS while the closed model corresponds to the crystal structure of *Gsp*DAH7PS

for which the missing atoms have been modelled. The CM domains are coloured in green, the linker regions in purple and the main DAH7PS barrels in blue. The substrate PEP and the allosteric inhibitor prephenate are shown in sticks in (A) and (B) respectively while the metal ions are represented as orange spheres. For each model, the inset represents a zoom on a particular region of the CM domain which is disorganised in the open form model and reorganised to accommodate prephenate in the closed form model. In these insets, the residues important for the binding of prephenate are displayed with sticks and prephenate is shown with grey sticks. **C.** Molecular surface representation of the closed (top) and open (bottom) models of *Gsp*DAH7PS. The active site of the DAH7PS moiety show up in the open form model and is identified with a red arrow which points on PEP (shown in sticks).

Comparison of these open and closed models sheds light on the possible structural changes associated with the binding of prephenate and mechanism of the allosteric inhibition of *Gsp*DAH7PS. In the absence of prephenate, there is little contact between the CM domain and the DAH7PS main barrels (7.6% solvent accessible surface area of tetramer buried, or 6% of the buried solvent accessible surface area of each of the CM dimers) and the structure is extended, with the entrance to the active sites of the DAH7PS domains accessible for substrate binding (Figure 5.4 A). In comparison, the structural model with prephenate bound (Figure 5.4 B) is far more compact, with the CM domains slightly bended to fit closely the DAH7PS main barrels and maintain close contact (15.6% of the solvent-accessible surface area of the tetrameric DAH7PS barrels buried by the CM domains), occluding in part the entrance to the active sites suggesting an explanation for the inhibitory activity of prephenate against the DAH7PS activity (Figure 5.4 C).

The CM domains of the open form model of *Gsp*DAH7PS present partially unfolded inter-helical loops and N-terminal regions in comparison with the model of the closed prephenate bound form for which the same region is reordered upon binding of prephenate. This change is especially apparent when morphing the two models (Movie 5.1 – Appendix A – Supplementary movies). This part of the CM domains includes residues identified as being

responsible for the catalytic activity of the CM domains and for the binding of prephenate, in particular, Ile34, Lys38, Asp47, Arg50 and Glu51 are the most affected. We can, therefore, postulate that the binding of chorismate or prephenate is accompanied by a reorganisation of the secondary structure of the CM domains, *viz.* the formation of longer helices. This reorganisation, together with the binding of the allosteric ligand itself, rigidify drastically the CM domains which find an energy minimum making up for this strain by leaning on the DAH7PS catalytic barrels.

5.3.3 Small angle X-ray scattering data confirm a conformational change upon addition of prephenate, consistent with the transition from an open to a closed conformation

Experimental SAXS data were obtained for *Gsp*DAH7PS in the absence and presence of prephenate in order to shed light on the conformation changes associated with allostery (Figure 5.5). Comparison of the SAXS profiles of *Gsp*DAH7PS in presence and absence of prephenate shows clear differences, consistent with a conformational change resulting from binding of prephenate (Figure 5.5A). This change is further confirmed as an increased compactness of the enzyme in the presence of an allosteric inhibitor by the smaller radius of gyration, derived from the Guinier plot, for the prephenate bound form (Rg = 34.8 ± 0.6 Å) than for the apo form (Rg = 36.5 ± 0.4 Å) (Figure 5.5A). The bell-shaped curve observed in the Kratky plot (Figure 5.5B) shows that both in the presence and absence of prephenate, the protein exhibits a folded shape without long random coils. It confirms that the CM domains remain folded even in the absence of prephenate, although the slightly more flattened curve for the prephenate-free sample may indicate an increased flexibility of the protein. The Porod-Debye plots (Figure 5.5C) present two different plateaux in the presence and absence of prephenate, suggesting that the protein

experiences a discrete conformational change upon ligand binding without any radical changes in its intrinsic flexibility [282,283]. This is supported by the $q^3 \bullet I(q)$ *vs.* q^3 plot (Figure 5.5D) that does not plateau as one would expect in the case of a highly flexible enzyme. The differences observed in the SAXS data both in the absence and presence of prephenate are consistent with the crystal structure and the models described above, that is a more compact and less flexible structure in the presence of prephenate is observed.

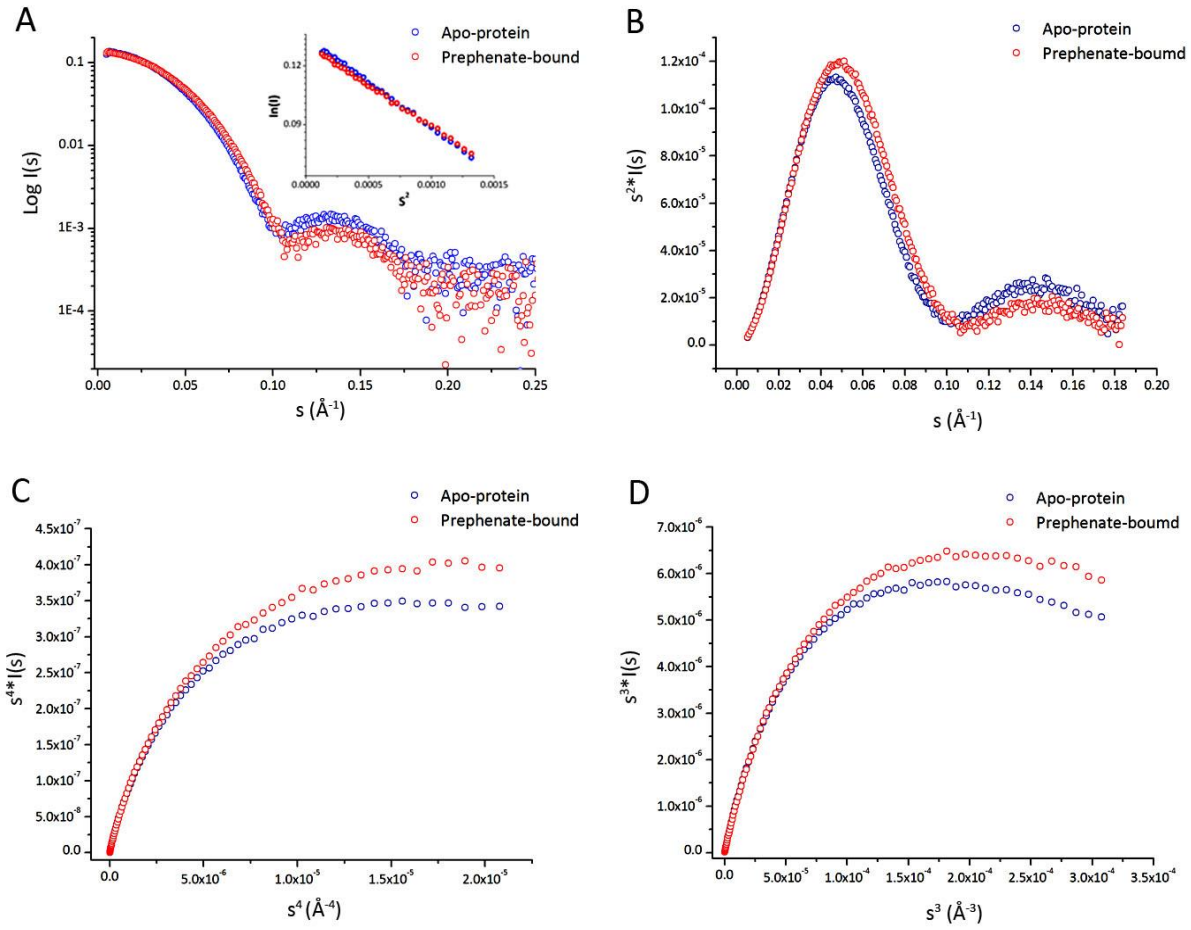


Figure 5.5. Analysis of the SAXS data for *Gsp*DAH7PS in the absence (blue) and presence (red) of allosteric prephenate. A. SAXS profiles (*log I(q) vs. q*) and Guinier plots (at top-right corner), **B**. Kratky plot ($q^2 \bullet I(q)$ *vs. q*), **C.** Porod-Debye plot ($q^4 \bullet I(q)$ *vs.* q^4) and **D.** SIBYLS plot ($q^3 \bullet I(q)$ *vs.* q^3), limited to the range of the SAXS data for which the Guinier linearity is observed.

Comparison of the models of *Gsp*DAH7PS in the presence and absence of prephenate with the experimental SAXS data was achieved by first modelling the non-cleavable His-tag on both models. The models were then tested against the experimental data with OLIGOMER [284] and the results are presented in Figure 5.6. The theoretical curves present a good fit with the experimental curves (χ^2 = 0.70 and 0.37 for the closed and open forms respectively). The experimental SAXS data were further fitted using a mixture of open and closed models of *Gsp*DAH7PS. In the absence of prephenate, the analysis shows a much better fit of mixed *Gsp*DAH7PS structure (χ^2 = 0.22) to the experimental profile than that of *Gsp*DAH7PS open form alone, predicting that ~73% protein adopts the open form structure, and ~27% protein adopts the closed form structure. There might be a conformational equilibrium between open and closed forms or, more likely, intermediate states between the open and closed form when the ligand is absent due to the motion of the CM domains, although the open structure or open-like structures are predominant. Whereas in the presence of prephenate, the best fit to the experimental profile (χ^2 = 0.37) is obtained with closed form structure alone, predicting that the closed structure or closed-like structures are the sole forms adopted by the prephenate bound protein. This indicates that the activity of *Gsp*DAH7PS can be totally inhibited by prephenate, which is consistent with the results of kinetic assays.

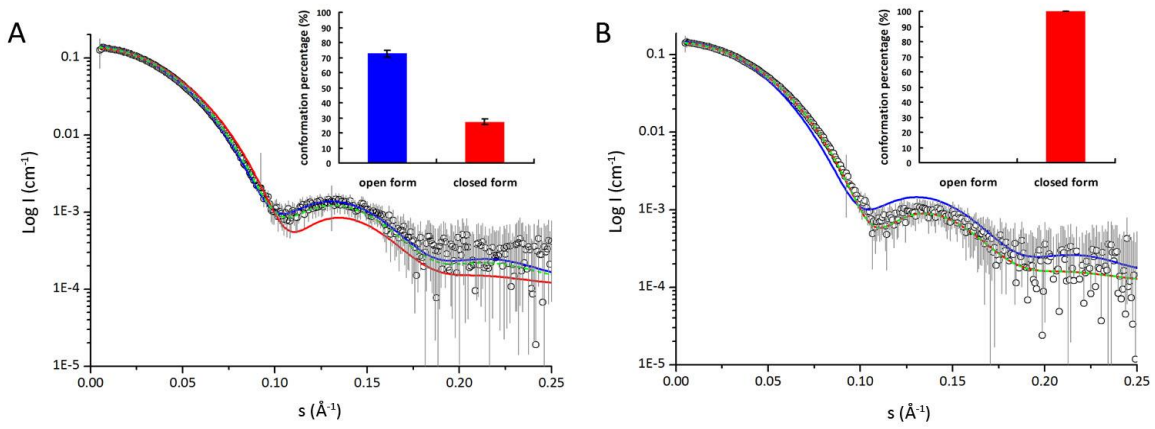


Figure 5.6. The distribution of open and closed conformations in GspDAH7PS with or without prephenate in solution. Theoretical scattering profiles were generated from models of both the prephenate-free open form, (Blue line) and the prephenate-bound closed form (red line) using OLIGOMER. These theoretical profiles were compared with the experimental GspDAH7PS SAXS data (open circles) in both the absence (**A**) and the presence (**B**) of prephenate. The experimental SAXS data were fitted to a distribution of both open and closed forms using OLIGOMER (green dashed line), and column charts show the percentages of open and closed forms GspDAH7PS adopted in the solution with and without prephenate.

Coupled with the crystal structures and homology models presented above, the experimental SAXS data clearly demonstrates a shift from open form to closed form in *Gsp*DAH7PS resulting from the binding of prephenate. The results from the analysis with OLIGOMER shows the existence of the closed form in prephenate-free *Gsp*DAH7PS, indicating that the effect of prephenate on *Gsp*DAH7PS may not induce the conformational change, but to stabilise the closed form in the conformational equilibrium.

5.4 Discussion

Based on these results, it seems that the CM domains can inactivate the enzyme by blocking the entrances to the active site, very much like the ACT domain of *Tma*DAH7PS. The SAXS data suggest that the CM domains remain tightly bound to the catalytic barrels in the presence of allosteric ligand, owing to the very good fit obtained for the data in the presence of inhibitor, using the closed form model.

In the absence of allosteric ligand, the CM domains remain at a distance from the catalytic barrels, permitting the substrates to enter the active site. However, the absence of

ligand also makes the CM domains more flexible and able to sample structures that resemble the closed form as indicated by the better fit obtained for the SAXS data by combining open and closed form structures. This is, again, similar to what has been observed for *Tma*DAH7PS (Chapter 4), and as the motions of the ACT domains affect the binding of E4P and the turnover number, the movements of the CM domains have similar effects on *Gsp*DAH7PS. This is illustrated in Table 4.3: k_{cat} is lower for the wild type enzyme than for a truncated mutant lacking its CM domains, while K_m^{E4P} is higher.

Table 5.1. Kinetic parameters for *Gsp*DAH7PS

	K_m^{PEP} (μM)	K_m^{E4P} (μM)	k_{cat} (s^{-1})
*Gsp*DAH7PS^{WT}	87 ± 9	95 ± 8	45 ± 4
*Gsp*DAH7PS^{truncated}	98 ± 11	62 ± 5	63 ± 5

However, in contrast to the ACT domains, the CM domains present only limited degrees of freedom because the dimerization of CM effectively leads to the CM domains being attached to the catalytic barrels by two linkers. However, it is not clear how this affects the activity of the enzyme when compared with *Tma*DAH7PS as on the one hand it can reduce the ability of the CM domains to move toward the active site in the absence of the ligand, but on the other hand, it also means that the enzyme is less likely to explore broadly open conformation that leaves the active site completely in unhindered.

Of particular interest as well is the partial folding of a random coil in the open form to generate longer helices that bind chorismate/prephenate. This raises, in particular, the question of whether the folded state of this region is sampled in the absence of ligand, or if it is the presence of chorismate or prephenate that triggers this folding. Moreover, how does the folding of this unfolded loop affect the motions of the CM, *i.e.* does the folding of this region initiate the closing of is the folded state only attainable once the CM domains cap the catalytic barrels?

Therefore, although X-ray crystal structure and homology modelling provides invaluable details on the allosterically inhibited and ligand-free state of the enzyme, and SAXS experiments enable to examine *Gsp*DAH7PS from a dynamical point of view, neither of these methods provide an atomic description of the protein motions nor provide any details on the transition between the two end states. Therefore, in order to answer the remaining questions about the allosteric mechanism of this enzyme, further investigations must be carried out.

5.5 Toward a better understanding of *Gsp*DAH7PS dynamics

In order, to gain more insight into the allosteric regulation mechanism of *Gsp*DAH7PS, MD simulations of this enzyme are currently in progress. However, at this stage, not enough data have been collected to conduct an in-depth analysis of the molecular motions associated with the allosteric regulation of the enzyme. Herein, the strategy employed will be presented and some details on the methodology used will be provided.

5.5.1 Classical MD simulations of GspDAH7PS

Starting from the closed and open models of *Gsp*DAH7PS described in sections 5.2.1 and 5.2.2 respectively, classical all-atom MD simulations were initiated. More specifically, three systems were prepared: the open form in the absence of ligand, the closed form in the presence of prephenate and the closed form in the absence of ligand. Simulations are carried out with NAMD [89] using CHARMM all-atom parameter set 22 with the CMAP correction [82,262]. Parameters and topology for prephenate were obtained from the CHARMM General

Force Field (CGenFF) [285] and were optimised using Force Field Toolkit (fftk) [286] and Gaussian09 [287]. After minimization, heating and equilibration, three runs were started for each one of the three systems, using different velocity distributions in each case. Each simulation is running until reaching 300 ns.

Once the simulations are finished, among other analysis methods, the resulting trajectories will be analysed in light of the experimental SAXS data. Indeed, each conformation generated by MD (*i.e.* each trajectory frame) for both the open and closed form in the presence of prephenate, can be converted into a PDB file, which can then be used to fit the experimental data. The resulting SAXS fits can then be analysed on their own or as part of a multi-conformational ensemble in solution. In this case, the multi-conformational scattering I(q) can be calculated by averaging the individual scattering of each N conformers, as shown in equation 5.1 [288]:

$$I(q) = \frac{1}{N}\sum_{j=1}^{N} I_j(q) \qquad (5.1)$$

One can either give the same weight to each conformer or automatically give more weight to a subset of conformations depending on the fitting using the MES programme [289]. I have written and tested the scripts to enable this analysis. This method should permit to confront experimental and computational data regarding the dynamics of *Gsp*DAH7PS in solution, to judge the comprehensiveness of the conformational landscape sampling by MD and to provide more information on the CM domain motions.

5.5.2 The use of Markov state models to study the transition between open and closed forms

In parallel to the classical MD approach, a large number of short simulations have been started in order to study the transition between open and closed form. The strategy in use here is partly inspired by the work of Silva and co-workers [111]. The idea is to first simulate the closing and opening of the CM domains using targeted MD (TMD). Then, multiple, regularly spaced conformations generated from the TMD runs are used to seed short classical all-atom MD simulations. These multiple short trajectories, at equilibrium, will allow to explore the energy landscape around the non-equilibrium path generated by TMD and can be used to create a Markov state model (MSM). This MSM should provide mechanistic and kinetics information on the open to closed and closed to open transitions and may be to shed light on the induced fit vs. conformational selection mechanism using, for example, an approach similar to that of Gu *et al.*[290], which is based on flux analysis [291] or to help assessing how well prephenate binds the representative structure of each metastable state along the path using molecular docking.

Regarding the practical implementation of this strategy, a theoretical truncated mutant of *Gsp*DAH7PS, which is only equipped of CM domains on one side of the protein (*i.e.* only one CM dimer), has been modelled. The reason for this choice is primarily to reduce the computational burden and speed up the calculations since the number of protein atoms, water molecules and ions are significantly lower than for the wild-type protein. In addition, by having only one dimeric CM domain, sampling of the conformations should be much more efficient and the construction of the MSM should be significantly simplified. Indeed, in the case of the wild-type enzyme, if one of the CM dimers for one frame has almost the same conformation than the CM dimer of another frame, but if the second dimer of each frame differs, the structures might not be clustered together (depending on the clustering parameters). However, this situation cannot arise in the case of the fictitious truncated mutant used here, and the two

structures would be grouped together, which makes, from the point of view of studying the transition between open and closed forms, more sense.

In addition, because in TMD simulations global conformational changes that greatly affect the RMSD occur prior to the local conformational changes, that have only limited effect on the RMSD, both forward and backwards TMD simulations need to be run, and the classical MD simulations to be started from both structures. Furthermore, the CM dimer of both the closed and open forms of *Gsp*DAH7PS do not present symmetrical sides: for the closed form one side is closer from the catalytic barrel than the other and for the open form one side is more unfolded than the other. Because of this, all the possible transition scenarios need to be explored as shown in Figure 5.7. The TMD trajectories a, b, c, d and their symmetrical equivalents represent the possible transitions from the open and closed form models, whereas the TMD trajectories e, f, g, h, I and j cover the scenarios in which the transitions occur from more symmetrical intermediate states in which the CM domain is in a more horizontal position.

For each end state structure, two classical MD simulations have been run for 50 ns. In addition, 9 equally distant structures are selected along each TMD trajectory. Each of these structures, after energy minimization, heating and equilibration, serves as a starting point for a 40 ns classical MD simulation. Simulations are conducted with the GPGPU version of AMBER [292]. Indeed, AMBER on a single GPU can reach performances similar to that obtained with hundreds of CPUs, depending on the type of hardware used, enabling to run efficiently a very large number of short simulations. At this stage, 20 ns for each structure have been simulated. Once 40 ns have been reached, a first MSM will be constructed in order to identify regions of the energy landscape that require additional sampling. Once this is achieved, a final MSM can be constructed.

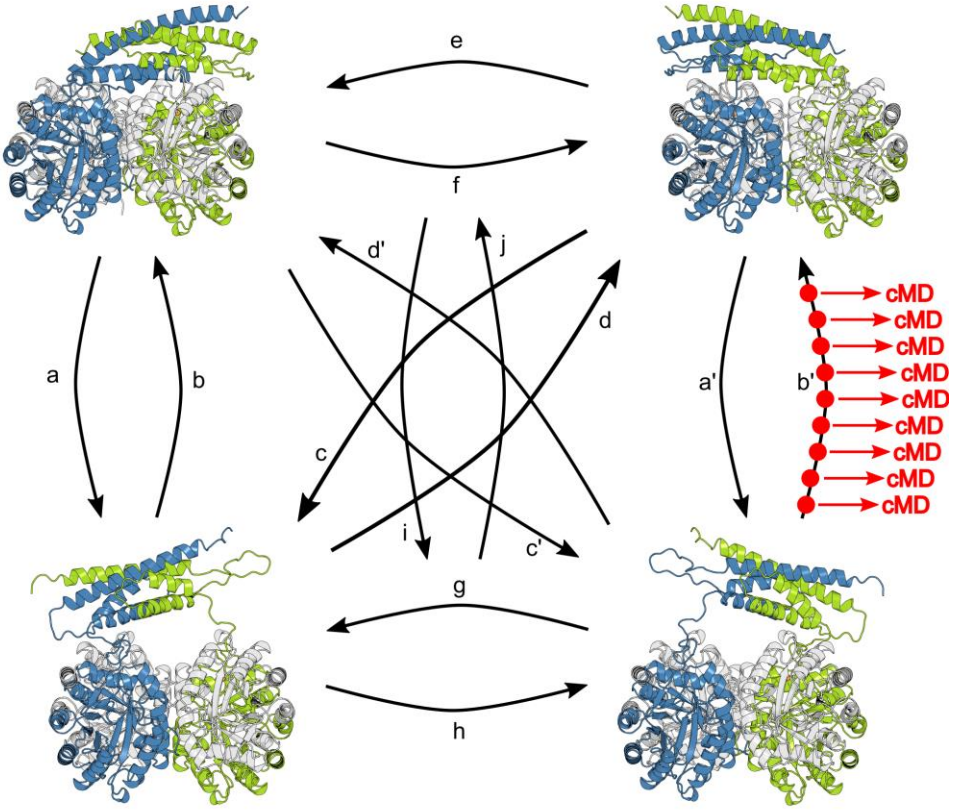


Figure 5.7. Schematic representation of the strategy adopted to sample the conformational transition between open and closed form. For each structure, chain A is in blue and chain B in green while truncated chains C and D (lacking the CM domain) are shown in grey. The black arrows represent the TMD trajectories and symmetrical trajectories are identified with a prime. For one of these trajectories, the structures selected to run short classical MD simulations are identified with red circles.

Chapter 6 – Toward the modelling of the first reaction step catalysed by DAH7PS

6.1 Introduction

DAH7PS catalyses the condensation of phosphoenolpyruvate (PEP) and D-erythrose 4-phosphate (E4P) to yield 3-deoxy-D-*arabino*-heptulusonate 7-phosphate (DAH7P). A number of kinetic, mutational and structural studies [179,183,293,294] have refined a plausible mechanism for this reaction that best agrees with the experimental evidences and is shown in Figure 6.1 [293,294]. The proposed mechanism predicts that first nucleophilic attack of PEP on the *re* face of the aldehyde moiety of E4P takes place, facilitated by the metal ion coordinating the oxygen of the carbonyl group of E4P. The resulting oxyanion is then protonated via an adjacent Lys residue leading to intermediate **1**. A water molecule, which is positioned on the *re* face of PEP, is then deprotonated by a Glu residue to attack the oxocarbenium ion leading to the tetrahedral intermediate **2**. Subsequent protonation of the phosphate group via a Lys residue facilitates the elimination of the phosphate yielding **3**. Deprotonation of the 2-keto moiety of **3** by the same Lys permits the restoration of its original protonation state and release of DAH7P. Glu145 and Lys99 are expected to be regenerated to their original state thanks to a water molecule.

Although the proposed mechanism is based on solid experimental evidence, to date no information exists on the energetic profile of the reaction nor on the role of dynamics in the catalysis. Such information can be retrieved from QM/MM modelling of the enzymatic reaction, which would generate detailed data on each step of the reaction and help to further refine or modify the proposed mechanism. Furthermore, QM/MM calculations could provide more insight on how the changes in dynamics associated with allosteric regulation could affect individual steps of the reaction, for example by destabilising key intermediates or increasing energy barriers. A QM/MM study of the related KDO8PS enzyme has been performed [295], but in this case, the enzyme does not need a metal ion to activate the nucleophilic attack between PEP and arabinose 5-phosphate (E5P) and the enzyme is more promiscuous, making any comparison of mechanism between KDO8PS and DAH7PS subject to caution.

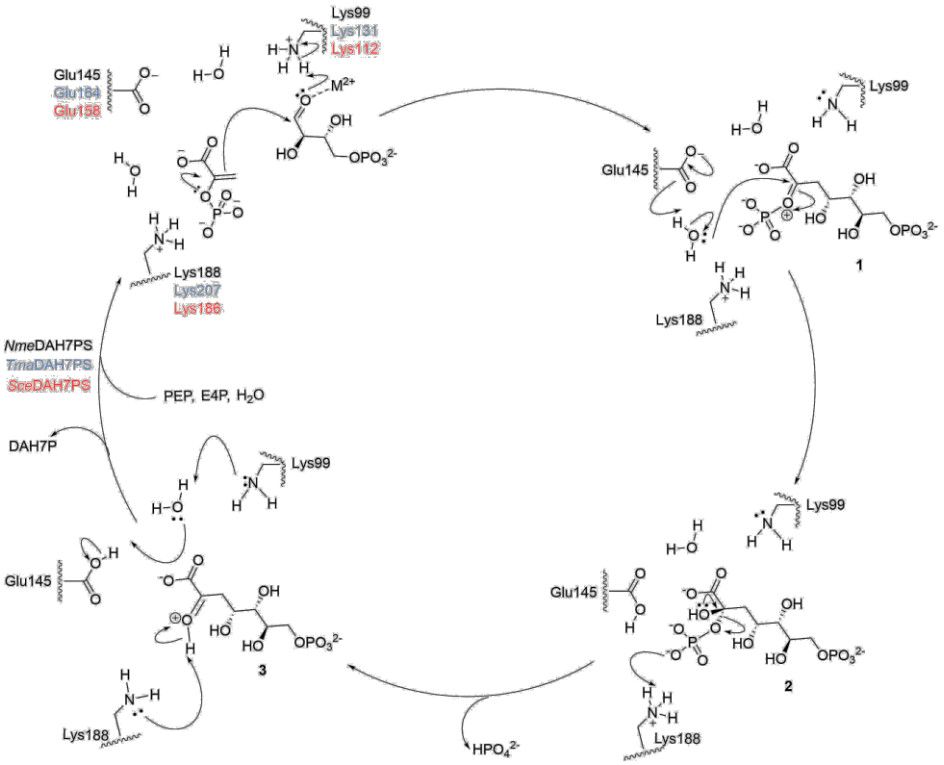


Figure 6.1. Postulated catalytic mechanism for DAH7PS. Adapted from references [293,294]. The residue numbering for NmeDAH7PS, TmaDAH7PS and SceDAH7PS is shown for the first reaction step in black, blue and red respectively.

In this chapter, the first attempts to model the first reaction step catalysed by DAH7PS, (i.e. the nucleophilic attack of PEP on E4P) and the subsequent protonation of the resulting oxyanion, are reported. In order to determine if the nucleophilic attack and proton transfer take place via a concerted mechanism or as a two stage process, both possibilities need be considered. As a first approach, QM/MM-MD implemented in AMBER 14 [84,296,297] was used to model this first reaction step. The objective of this approach was to enable a systematic exploration of the free energy landscape of the reaction, while the error from the semi-empirical (SQM) method used could be assessed later by comparing the energy of the key species obtained in gas phase with SQM and with a more accurate QM method. However, the SQM methods implemented in AMBER were unable to model correctly the phosphate groups and the transition metal, and an adiabatic mapping approach was considered instead. Extensive testing and optimisation of the adiabatic mapping was performed but unfortunately, modelling of the reaction could not be achieved. Nevertheless, this work paves the way to the modelling of the first reaction step catalysed by DAH7PS.

6.2 Tentative modelling of the first reaction step of DAH7PS from *T. maritima* using a QM/MM-MD approach

6.2.1 Set-up of the simulations

The structure of *Tma*DAH7PS in its open form (PDB 1RZM) was used for the modelling as it has been co-crystallised with a metal ion (Cd^{2+}) and substrates PEP and E4P. The choice of the semi-empirical method is dictated by the presence of a metal ion and of the phosphate groups. AM1/d-PhoT includes the parameters for Zn^{2+} and has been developed for phosphates,

thus this method is the best choice *a priori*. The metal ion was replaced with Zn^{2+}, as it is known to adequately replace Cd^{2+} for catalysis, and the reaction was modelled in chain B since the position and orientation of both PEP and E4P are reasonable starting points to model the first reaction step. Indeed, in chain A, the carbonyl gorup of E4P is not modelled in a reactive conformation in the original structure. MM force field parameters and topology for E4P and PEP were obtained using the SwissParam server [272] and the parameters for the Zn^{2+} ion were those included in the CHARMM all-atom parameter set 22 with the CMAP correction [82,262].

The missing residue Glu261 was added with COOT [258] and the position of the side chain of one Met211 was changed to avoid steric clashes. The protonation state of each of the ionisable residues (except for histidine) was calculated with PROPKA 3.1 [261], averaging the results per chain. All residues were predicted to exhibit a standard protonation state except for Glu164, which is predicted to be protonated (pK_a = 8.2) and thus modelled accordingly. This deviation does not come has a surprise as Glu164 faces the negatively charged phosphate group of PEP, favouring a neutral species to avoid repulsive charge interactions. The protonation state of each histidine was calculated using MCCE [260] and modelled as follows: His29, His77 and His272 were singly protonated on Nδ1, His263, His293 and His300 were singly protonated on Nε2 and His115 was doubly protonated (positively charged). The protein was first solvated with a 3 Å explicit TIP3P solvent layer using Grubmüller's SOLVATE program [298] and then the resulting system was solvated in an explicit TIP3P water box of 110 × 110 × 115 Å. Na^+ and Cl^- ions were added with a minimum distance of 5 Å from the enzyme and from each other to neutralise the net charge of the system and reach an NaCl concentration of 0.1 M.

Simulations were carried out with AMBER 14 [292] using the CHARMM all-atom parameter set 22 with the CMAP correction [82,262], which is accurately represented in

AMBER owing to CHAMBER [299]. The particle mesh Ewald (PME) method was used to calculate electrostatic interactions and Van der Waals cut-off was set up at 12 Å. The simulations were conducted with a 2 fs time steps with the SHAKE algorithm used to constraint all bonds involving hydrogen atoms. The initial structure was first minimised with the conjugate gradient energy minimization method for 1000 steps with a harmonic restraint of 100 kcal/mol/Å^2 on all atoms except hydrogens, then for another 1000 steps without restraining the water molecules and ions. For 5000 steps the harmonic restraint was decreased to 50 kcal/mol/Å^2 and applied only to the backbone atoms and the metal ion. A final minimisation of 5000 steps was performed with a 5 kcal/mol/Å^2 restraint on alpha carbon and metal ions. The system was then heated from 0 to 333.15 K over 25000 steps, keeping the same harmonic restraint. The system was then equilibrated in the NPT ensemble, using Langevin dynamics to keep the temperature at 333.15 K and the Berendsen barostat to keep a constant pressure of 1 atm, with the restraint on alpha carbon and zinc ions decreased stepwise from 5 to 0 kcal/mol/Å^2 over 25000 steps and the systems equilibrated without any restraint for 500000 steps.

In order to test the stability of the substrates within the active site, 10 ns of unrestrained MD simulation were performed, starting from the equilibrated structure. During this run, major motions of PEP and E4P were observed, resulting in the carbonyl of E4P presenting the wrong face to PEP or being too far away to model the reaction. In addition, Lys131 (presumed to be responsible for the protonation of the oxyanion) was shifted away from E4P, and the residues coordinating Zn^{2+} moved away from the metal ion, leaving it partially uncoordinated. These observations are likely mainly a consequence of the limitation of the force field, in particular the parameters used to model E4P, PEP and Zn^{2+}.

In order to avoid these problems, the same minimisation, heating and equilibration protocol was applied, but this time a number of restraints were applied to the key distances throughout the whole procedure and kept for a 1 ns MD simulation. More precisely, restraints of 50 kcal/mol/Å^2 were applied to the distance between the metal ion and the oxygen of the aldehyde of E4P, between the metal ion and the nucleophilic carbon of the PEP double bond, and between this carbon and the electrophilic carbon of E4P aldehyde function. Despite these new constraints, significant deviations in the position of the substrates were still observed. Therefore, a shift to a QM description of the active site immediately after MM minimization was chosen in order to avoid these deviations.

Using the AM1/d-PhoT/CHARMM-22 level of theory, minimization of the structure for a minimal QM region consisting in PEP, E4P (but not its phosphate group), Zn^{2+}, and part of the side chain of Lys131 was performed. After QM minimisation of the MM-minimised system, a twist of the carbonyl of E4P from its ideal position with respect to PEP was observed (Figure 6.2).

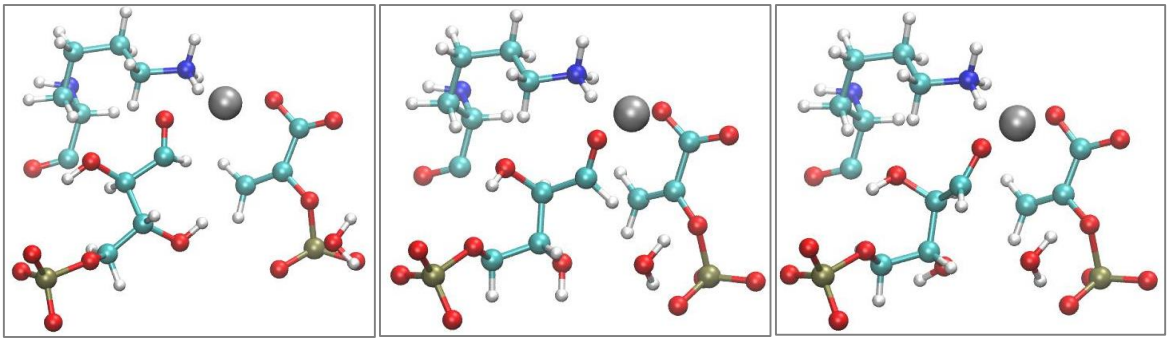


Figure 6.2. Relative positions of PEP, E4P, Zn^{2+} and Lys131. Left: initial solvated crystal structure. Central: MM-minimised (CHARMM-22). Right: QM/MM-minimised (AM1/d-PhoT/CHARMM-22).

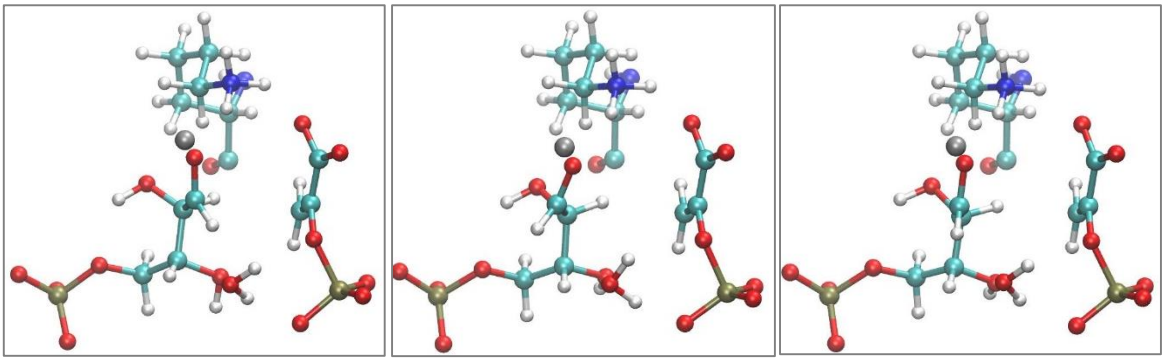


Figure 6.3. Relative positions of PEP, E4P, Zn^{2+}, Lys131 and a water molecule. Left: initial MM-minimised structure. Central: QM/MM-minimised without adding the water molecule to the QM region. Right: QM/MM-minimised with the water in the QM region.

Interestingly, a water molecule (which is not found crystallographically, but comes from the solvation shell) moves in the direction of the carbonyl during the MM minimization and remains at an interacting distance of PEP and E4P during a 10 ns MD simulation using MM force field and restraints. The inclusion of this water molecule within the QM region led to a reduction of the twist of the carbonyl group with the carbonyl being parallel to the double bond of PEP (Figure 6.3), presenting a better positioning for the attack of PEP with a better angle for the HOMO-LUMO overlap. Furthermore, in this configuration, the carbonyl is oriented such that one of the lone pairs of the oxygen is directed toward Lys131 and the other toward Zn^{2+} with the carbonyl carbon at a reasonable distance from the double bond for the modelling of the bond formation (Figure 6.4).

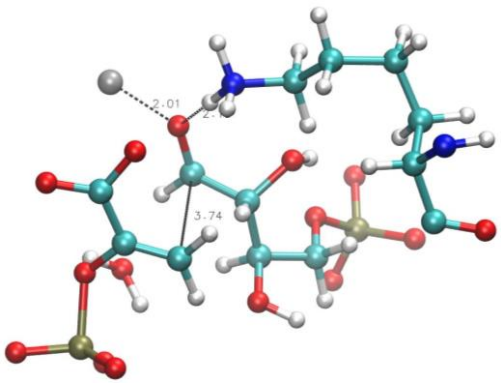


Figure 6.4. PEP, E4P, Zn^{2+}, Lys131 and a water molecule after QM minimisation.

A number of QM/MM-minimisation tests were run with an increasing number of atoms in the QM region: residues coordinating the metal ion, as well as those interacting with PEP. These tests did not lead to important changes in terms of geometry and distances. Thus, the minimal QM region, including the water molecule described above represents a reasonable starting point for the calculations from a structural perspective.

6.2.2 Test QM/MM-MD modelling

A tentative unrestrained QM/MM-MD simulation of 100 ps using the AM1/d-PhoT/CHARMM-22 level with the minimal QM region including the water molecule (see section 6.2.1 and Figure 6.3) led to the cleavage of the O-P bond of PEP after about 25 ps and release of the phosphoryl group (Figure 6.5). As a result, the remaining part of PEP seems to rearrange as a carbanion that coordinates Zn^{2+}. The addition of the lysine and arginine residues that bind the phosphate group of PEP (Arg237, Lys207 and Arg126) to the QM region does not change the outcome of the simulation.

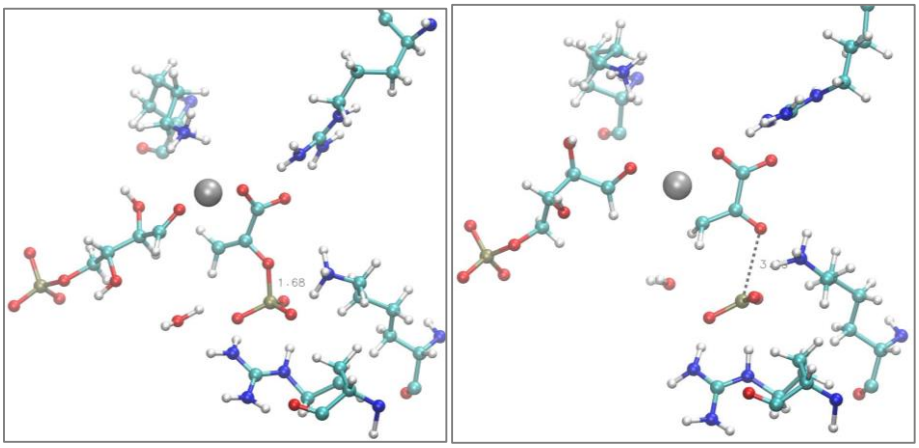


Figure 6.5. PEP, E4P, Zn2+, Lys131, Arg237, Lys207, Arg126 and a water molecule. Left, after MM equilibration, before QM/MM calculations. Right, after about 25 ps of MD using a QM/MM potential.

This is in direct contradiction with the ^{18}O labelling experiments that demonstrated a C-O bond cleavage mechanism rather than an O-P cleavage mechanism operates for DAH7PS [300] as well as in the related KDO8PS [301] and NeuB [302]. Observation of the O-P bond cleavage is, therefore, most likely due to the limitations of AM1/d-PhoT and is probably related to the fact that this SQM method has been mainly parameterised based on O-P bond formation and cleavage [303,304]. The alternative SQM methods (PM6, PM3 and AM1) tested worsened the results and, therefore, AM1/d-PhoT was used, but with a strong (400 kcal/mol/Å^2) harmonic restraint on the P-O bond in addition to a 100 kcal/mol/Å^2 restraint used to restraint Lys131 at a reasonable distance from the carbonyl oxygen of E4P.

6.2.3 Modelling the bond formation between PEP and E4P

Using the two restraints mentioned above, a test umbrella sampling at the AM1/d-PhoT/CHARMM-22 level of theory and the QM region consisting of PEP, E4P (without its phosphate group), Zn^{2+}, part of Lys131 and the water molecule was performed using the distance between the C1 of E4P and the C3 of PEP as reaction coordinate, from 3.7 Å to 1.4 Å. A window size of 0.1 Å with a simulation time of 2 ps for each window and a force constant of 200 kcal/mol/Å^2 was used. Umbrella sampling enables to sample exhaustively the conformations along the reaction coordinates, including the transition state, and calculate the free energy profile along these reaction coordinates.

Plotting the region sampled during the umbrella sampling shows a major gap between approx. 2.2 and 1.8 Å (Figure 6.6 - left), which is confirmed by the histogram plot (Figure 6.6 - right). This gap corresponds to the region around the transition state (TS) and a stronger restraint in this region will be needed for a correct sampling.

Interestingly, monitoring of the P-O distance in PEP over the course of the simulation reveals a sudden increase of the bond length upon the crossing of the transition state, and this despite the very strong harmonic restraint on this bond (Figure 6.7). Therefore, it seems that any electronic changes occurring at a reasonable distance of this phosphate might reinforce the tendency of this P-O bond to cleave and consequently affect the energy calculation of the corresponding reaction step. This will be particularly true for the third step of the DAH7PS reaction (i.e. cleavage of the C-O bond) since the willingness of P and O to be apart from one another will considerably affect the energy required to cleave the C-O bond. Monitoring of the distance between the N of Lys131 and the oxygen of the carbonyl of E4P does not reveal such a dramatic change upon the crossing of the TS, suggesting that in our model, the proton transfer is not spontaneous.

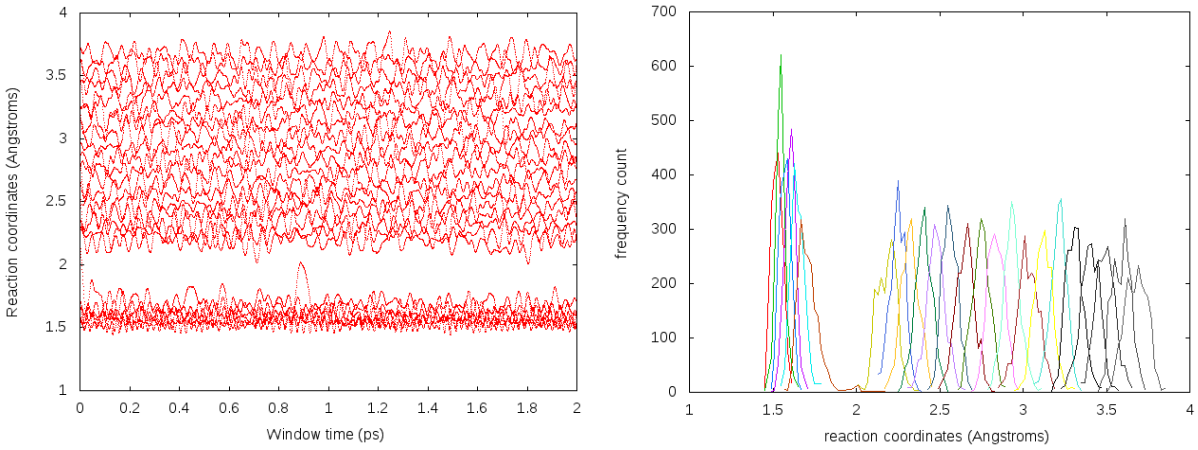


Figure 6.6. Umbrella sampling. Left, region sampled during each umbrella sampling window. The gap indicates the region around the transition state. Right, histogram plot for each umbrella sampling window

Using the weighted histogram analysis method (WHAM) [305], the relative free energy of the reaction along the reaction coordinate can be assessed and is plotted in Figure 6.8. It predicts a transition state at a reaction coordinate of 1.9 Å with a free energy barrier of about

20 kcal/mol, while the optimal bond length in the product is 1.6 Å. The product is of higher energy than the reagents (free energy difference of about 10 kcal/mol), which is not surprising since this product corresponds to an unprotonated intermediate. These results represent a very crude approximation given the flaws of AM1/d-PhoT for the phosphate of PEP and because the umbrella sampling simulation was set up as a test run (TS was not sampled correctly, short simulation time was selected for each window). The trajectory of this simulation is presented in Movie 6.1 (Appendix A – Supplementary movies).

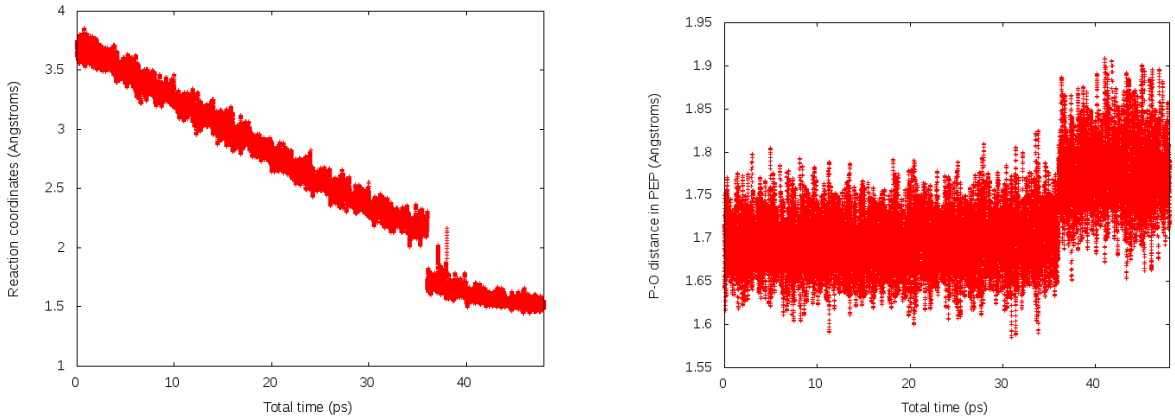


Figure 6.7. Monitoring of the restraints during the umbrella sampling simulation. 24 windows x 2 ps = 48 ps. left, the distance between C3 of PEP and C1 of E4P. Right, P-O bond length in PEP.

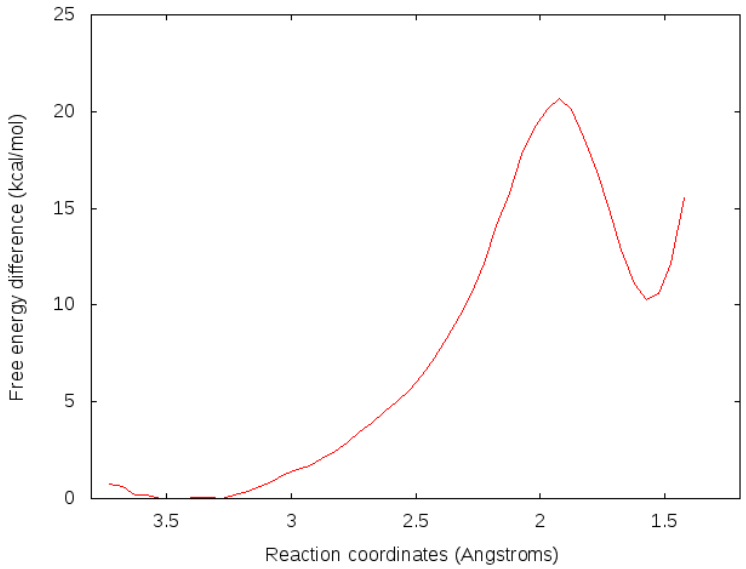


Figure 6.8. Relative free energy along the reaction coordinates. Free energy assessed with WHAM at a convergence tolerance of 10-9.

6.2.4 Conclusion on the QM/MM-MD approach

These results highlight some of the weaknesses of AM1/d-PhoT to model, at least, the first step of the reaction. However, since other common SQM methods (AM1, PM3 and PM6) are problematic for describing the chemical species involved and because SCC-DFTB (DFTB2) parameters for a combination of phosphate and transition metals have not been developed (and DFTB3 is not yet implemented in AMBER), the SQM approach was put aside and adiabatic mapping in conjunction with a higher level QM method was used instead (see section 6.3).

6.3 Tentative modelling of the first reaction step of DAH7PS from *T. maritima* using a DFT QM/MM approach

6.3.1 Set-up of the simulations

Following the limitations of AM1/d-PhoT to model the first step of the reaction as discussed in section 6.3, an approach based on using DFT instead of SQM was tested. A number of modifications have been made when compared with the set-up of the QM/MM-MD simulations in Amber. Firstly, Cys102, which coordinates manganese, is now modelled as being unprotonated, which is more consistent with the behaviour of Cys residues binding metal ions [306]. In addition, PROPKA predicts Glu164 to be protonated in the presence of PEP as

the pK_a averaged over the 4 chains is equal to 8.3 (no major differences between chains) and the protonotaion states are chosen for pH = 7.5. This protonation state is mainly due to the contribution of the phosphate group of PEP, which is at an average distance of 4.4 Å and Glu164, is predicted to be unprotonated in the absence of PEP. For the QM/MM-MD simulations, Glu164 was chosen to be protonated. However, based on the suggested mechanism described in Figure 6.1, unprotonated Glu164 is involved in the second step of the reaction and should deprotonate a water molecule that will attack the intermediate from the first reaction step. Moreover, a water molecule, conserved in many DAH7PS structures, is found between Glu164 and PEP and it is believed to be the one that attacks the intermediate in step 2 (WAT1 in reference [294]). Therefore, it seems reasonable to model Glu164 as unprotonated, so it may readily capture a proton from the adjacent water molecule for the next reaction step. Glu164 was therefore kept unprotonated for this simulation.

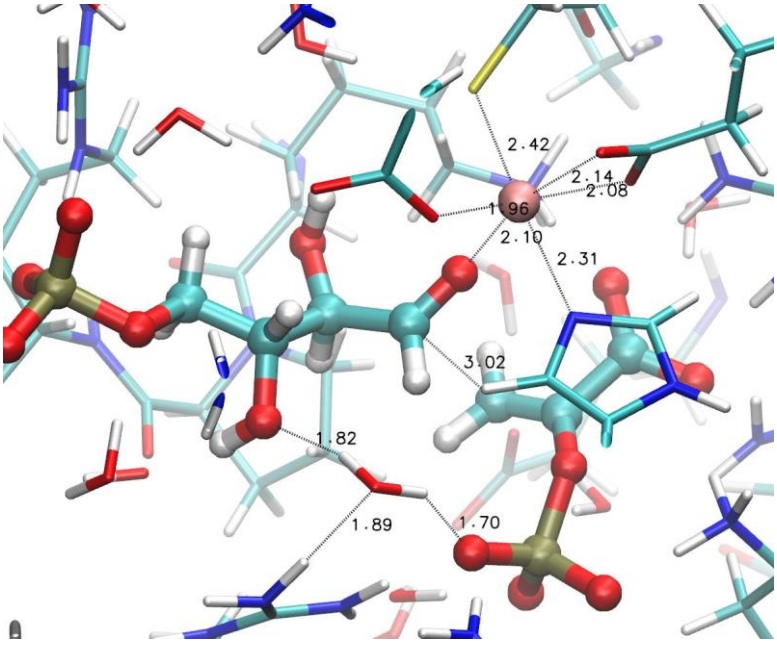


Figure 6.9. The first snapshot chosen for the DFT-QM/MM calculations.

In addition, manganese is modelled instead of Zn^{2+}. Previously, zinc was chosen because parameters were available for AM1/d (but not manganese). This does raise the question of the multiplicity that should be used in the simulations. After careful consideration, it was decided to first model the high spin state of Mn^{2+} (5/2) based on both experimental [307,308] and computational evidence. Indeed, in a recent QM/MM study by Thiel and co-workers [309], the energy of the sextet, quartet and doublet were compared and it appeared that the sextet was the only accessible state under mild conditions.

After minimization, heating and equilibration at the MM level (in AMBER) an equilibrated conformation ("snapshot") was selected based on the distances and geometry of PEP, E4P and Mn^{2+}. The snapshot was chosen so the distance between C1 of PEP and C1 of E4P is short, the carbonyl E4P presents its *re* face to PEP and the oxygen of the carbonyl group coordinates Mn^{2+}, as shown in Figure 6.9. During the MD equilibration, a harmonic restraint was applied to the distance between the carbonyl of E4P, C1 of PEP and Mn^{2+} as well as on all atoms of PEP, E4P, and Mn^{2+}. Removal of these restraints or setting up a spring constant too low inevitably leads to a twist of the carbonyl of E4P which ends up presenting its *si* face to PEP, which in turn would lead to the formation of the wrong stereocentre at this carbon in both the reaction intermediate and DAH7P product.

6.3.2 QM/MM simulations

After setup, MM equilibration and selection of a snapshot, QM/MM simulations were started with ChemShell [310], using DL-POLY [311] for the MM part and ORCA [312] for the QM part. A geometry optimization procedure was started using the B3LYP/CHARMM22 level of theory and using the def2-SVP basis set since it has been used with manganese

successfully before [309]. The entire system coming from MD equilibration (Figure 6.10 left) was first minimised for 10000 MM steps and the protein was truncated for computational efficiency so only the regulatory domain of chain A is kept, the others are cut at residue 71. A sphere of 25 Å of water centred on PEP, E4P and Mn^{2+} of chain B is kept as well as a solvation shell of 4 Å around the rest of protein. Ions present in the solvation shell are also conserved and a number of non-solvated ions are kept in order to neutralise the system (Figure 6.10 right). The active region (atoms allowed to move) was selected to be within a sphere of 5 Å radius centred on PEP, E4P and Mn^{2+} of chain B. The geometry optimisation was then started with the BP86 functional and the much smaller 3-21+G* basis set for the first optimisation round. Indeed, although it is expected to yield less accurate energies, BP86 is faster than B3LYP and geometries are expected to be comparable [313]. B3LYP with the def2-SVP basis set could then be used for a second round of optimisation.

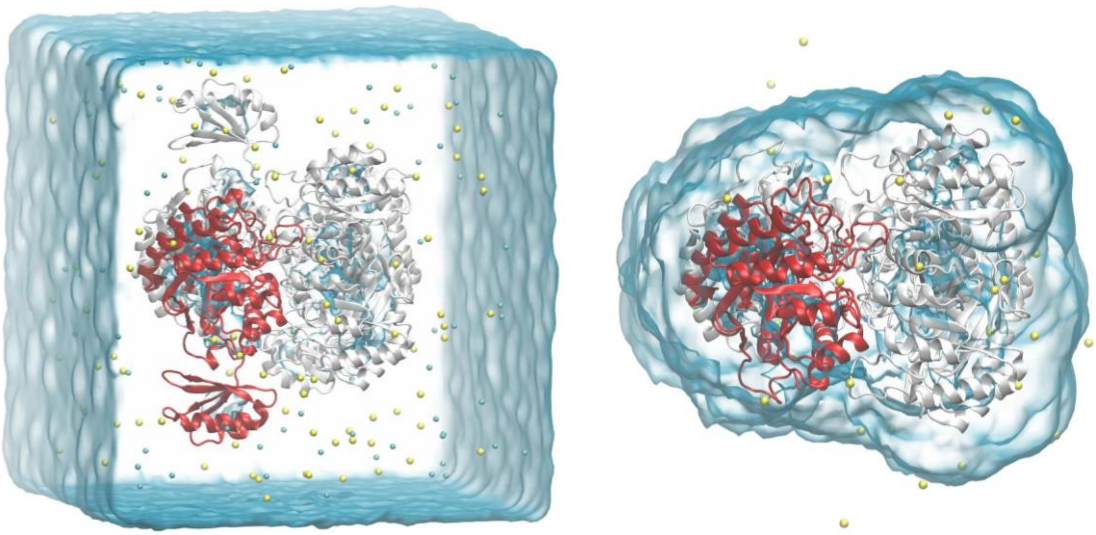


Figure 6.10. System before (left) and after (right) truncation

Three different QM regions were optimised for 500 cycles using the BP86 functional paired with the 3-21+G* basis set for the QM region.:

- QM1: Mn2+, PEP, E4P (without its phosphate) and Lys131 (cut between Cβ and Cγ).

- QM2: QM1 + water molecule interacting with both a hydroxyl of E4P and the phosphate of PEP identified as being important in the QM/MM-MD simulations.
- QM3: QM2 + Mn^{2+} coordinating residues (Cys102, His272, Glu298 and Asp309)

Convergence was not reached for all three systems after 500 steps. For QM1, very rapidly the carbonyl of E4P is twisted as it is during MD in the absence of restraints (Movie 6.2 – Appendix A – Supplementary movies), additionally, PEP is slightly flipped on its side moving significantly from its starting position. For QM2, the same observations stand but in addition, the water molecule included in the QM region is spontaneously deprotonated by the twisted carbonyl of E4P (the proton appears also to be partially shared with the second hydroxyl of E4P), releasing a free hydroxide (Movie 6.3 – Appendix A – Supplementary movies). For QM3, the addition of the residues coordinating the metal ion results in a partial twist of E4P and the metal ion partially leaving its coordinating environment and His272 interacting with the carbonyl of PEP (Movie 6.4 – Appendix A – Supplementary movies). Mn^{2+} loses its pyramidal coordination geometry as shown in Figure 6.11. In order to see if these surprising results are due to the use of a level of theory that does not accurately account for the electrons of manganese, calculations for QM1 were also performed with the B3LYP/def2-SVP level of theory, but led to the same results. Using a different snapshot and performing the MM minimization using the same restraints as for the MD did not change the results of the simulations.

The effect of having Cys102 protonated does not change the outcome of the simulation neither does the replacement of Mn^{2+} with Zn^{2+}. Indeed, in both cases, if the metal coordinating residues are not included in the QM region, the carbonyl of E4P flips to present its *si* face to PEP and if the residues coordinating the metal ion are included, then again in both cases the metal ion leaves its coordinating residues (Movie 6.5 – Appendix A – Supplementary movies).

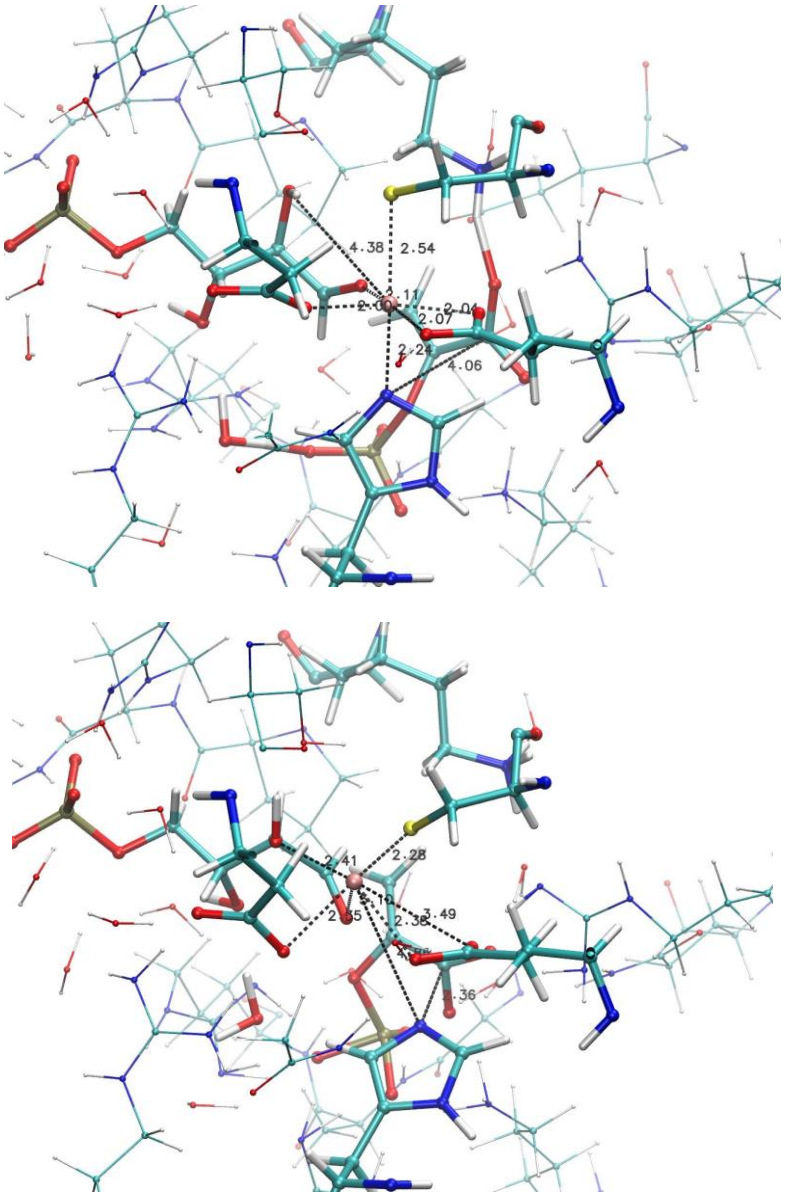


Figure 6.11. Coordination of manganese before (top) and after (bottom) QM/MM optimisation.

The observed behaviours in the calculations might arise from the original position of the active site residues and substrates, in particular, the carbonyl of E4P and the metal coordinating residues. One limitation of the current starting point is the position of E4P as the only clearly interpretable electron density in this structure is for the phosphate functionality (PDB 1RZM). Therefore, another structure might provide a better starting point of the simulations.

6.4 Tentative modelling of the first reaction step of DAH7PS from *S. cerevisiae* using a DFT QM/MM approach

Looking back at the solved structures of DAH7PS and the associated density, the structure of *S. cerevisiae* DAH7PS (*Sce*DAH7PS) in complex with PEP, Co^{2+} and glycerol-3-phosphate (G3P) (PDB 1OF8) represents an interesting starting point. G3P corresponds to E4P without the carbonyl group. The density around G3P is much better than the density of E4P in 1RZM. Modelling of E4P from G3P is straightforward by replacing the water molecule coordinating the metal ion with the oxygen of the carbonyl of E4P, yielding a good starting geometry for modelling the reaction. Moreover, the resolution of 1OF8 (1.5 Å), is the best obtained for a DAH7PS enzyme so far, and it gives very fine details of the position of the coordinating residues of Co^{2+}. Finally, the dimer can be modelled instead of the tetramer as it has been shown that the dimer is elementary biological unit in type Iα DAH7PS [248], which will reduce the number of atoms simulated in the MM region. After careful consideration, the decision was made to redo the calculation with *Sce*DAH7PS.

6.4.1 QM modelling of the metal centre

In order to develop effective QM/MM geometry optimisation protocols and to check if the metal ion can remain in position, pure QM geometry optimisations were performed with ORCA directly on the crystal structure of the metal centre. Modelling of the coordinating residues histidine, aspartate and glutamate, and cysteine was achieved using imidazole, acetic

acid and hydrogen sulfide respectively (H_2S has been shown to model the acidity of cysteine more accurately than other model compounds such as CH_3SH or CH_3CH_2SH [314]). After some tests, it appeared that the SCF convergence can be reached relatively rapidly by first using the BP86 functional with dispersion correction D3BJ and the small basis set def2-SV(P) using the RI-J approximation for the Coulomb integrals to speed up the calculation and the associated DeMon/J Coulomb fitting auxiliary basis set to approximate the charge distribution, and using the flags looseSCF and VerySlowConv. Once the geometry is optimised with this method, the B3LYP functional can be used with dispersion correction D3BJ, the larger basis set def2-TZVP and the RIJCOSX approximation (combined RI-J approximation for Coulomb integrals and COSX numerical integration for HF exchange). Geometry optimised structure of the metal centre using this process is shown in Figure 6.12. In both cases (HS^- or H_2S), the metal ion does not "escape", although the protonated form leads to a geometry closer to the octahedral geometry observed in the crystal structure.

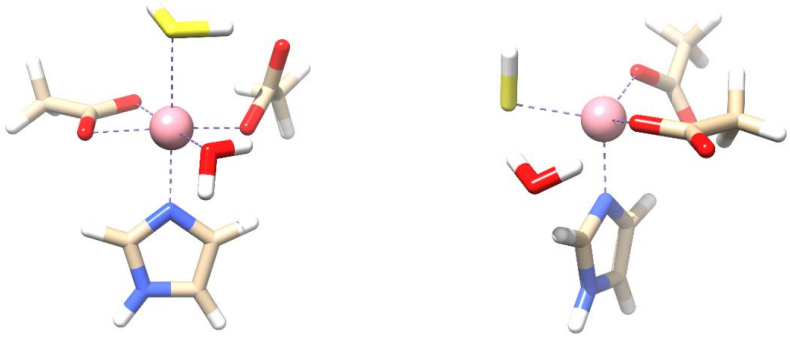


Figure 6.12. Optimised geometry of the metal centre using the method describe in the main text. Left: protonated cysteine model. Right unprotonated cysteine model.

In order to assess the effect of the electrostatic environment, geometry optimisations using a continuum solvent model were started from the gas phase optimised geometries using the SMD [315] water model. Optimisation converged rapidly with no major differences from the gas phase geometries. To further assess the impact of the environment, especially around

the coordinating cysteine, additional residues were included for the gas phase geometry optimisation. Both for the protonated and unprotonated cysteines, geometry optimisations at the BP86 def2-SV(P) level of theory did not lead to the escape of the metal ion (Figure 6.13). Deviations of the residues from their initial positions were observed and interestingly, the optimised geometries do not present an octahedral arrangement of the metal binding residues.

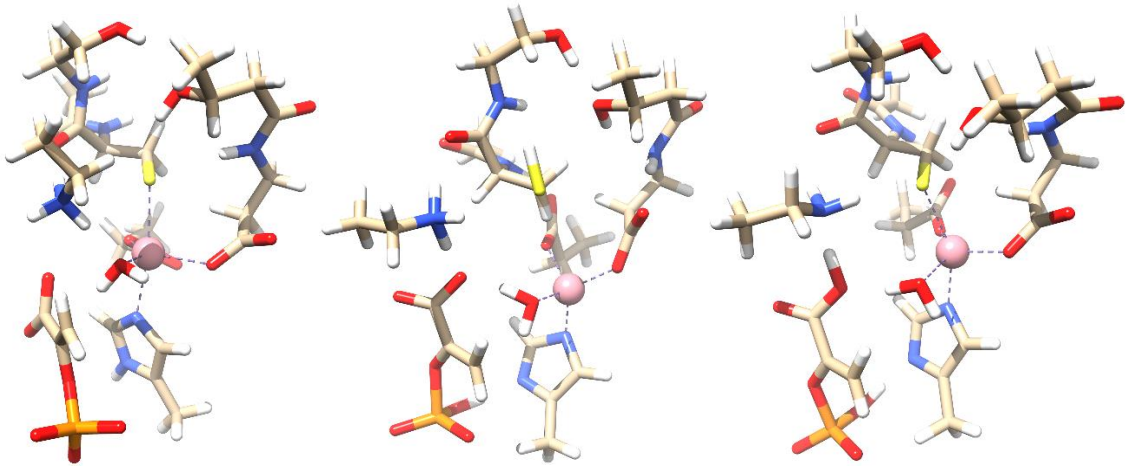


Figure 6.13. Geometry optimisation of the metal ion, coordinating residues, PEP and a few neighbouring residues at the BP86 def2-SV(P) level of theory. Left: starting geometry. Centre: optimised geometry with protonated cysteine. Right: optimised geometry with unprotonated cysteine.

6.4.2 Set-up of the simulations

Prior to starting the QM/MM calculations, E4P was modelled in COOT [258] based on the structure of G3P and the water molecule coordinating the metal ion (Figure 6.14). Missing residues (N-terminal tail and some residues of the $\beta_8\alpha_8$ loop) and missing side chains were modelled with MODELLER [259] while keeping all the other residues fixed. The best model was selected on the basis of the DOPE and MOLPDF scores given by MODELLER as well as on visual inspection. The structure of the best model was checked using the WHAT IF server

[316], the protonation state of the ionisable residues determined with PROPKA 3.1 [261] except for histidine residues for which the protonation state and position of the proton in neutral species were calculated with MCCE [260].

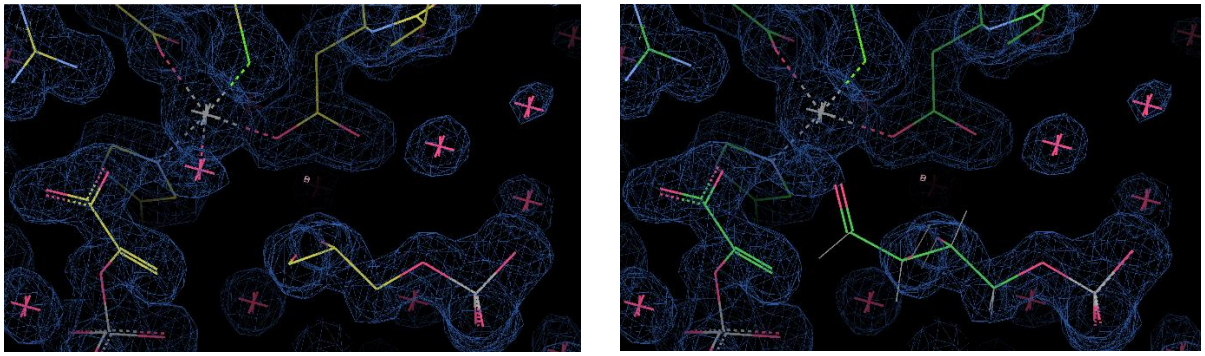


Figure 6.14. Modelling of E4P in COOT. Left: original structure and the associated electronic density, G3P is on the bottom right, PEP on the left and the metal centre on the top left corner. Right: E4P modelled in place of G3P and the water molecule.

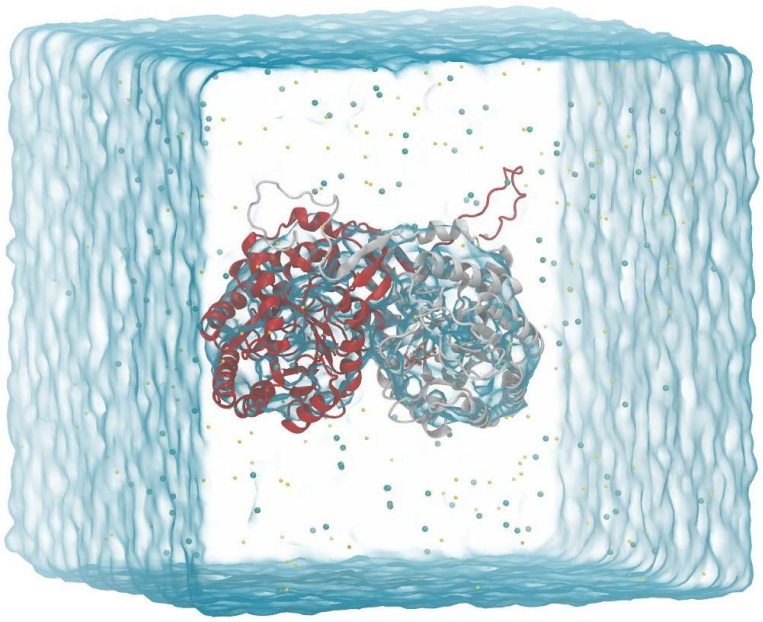


Figure 6.15: Dimer of *Sce*DAH7PS solvated and neutralised with Na^+ Cl^- at a concentration of 0.15 M.

Parameters for Co^{2+} were taken from the 'CM' set of parameters developed by Merz and co-workers [317]. Parameters and topology for E4P and PEP were calculated using the

MATCH server [318]. Crystallographic water molecules were conserved and hydrogen atoms were added to the structure using PSFGEN. Helmut Grubmüller's SOLVATE programme [298] was used to first generate a solvent shell around the protein and the system was then solvated in a relatively large solvent box (120 × 120 × 130 Å) to allow for the potential motions of the long N-terminal tails. Na^+ and Cl^- were used to neutralise the system and then added to reach a concentration of 0.15 M (Figure 6.15).

For the whole MM minimization and equilibration (performed with NAMD), harmonic restraints were applied to the distances and angles between the metal ion and its coordinating residues, as well as between PEP and E4P. Indeed, initial tests showed that without these restraints, the coordinating residues, the metal ions and the substrates deviate significantly from the crystal structure and might reach positions that are not optimal for the reaction and coordination of the metal ion. The simulations were conducted using a 2 fs time steps in conjunction with the SHAKE algorithm. The particle mesh Ewald method was used to calculate electrostatic interactions and Van der Waals cut-off was set to 12 Å.

All atoms were constraint except water molecules, ions and hydrogens atoms and the system was minimised for 5000 steps using the conjugate gradient energy minimization method, then the side chains of the protein, as well as the backbone of the N-terminal tail, were released and the system minimised for another 10000 steps. Finally, the whole system (except coordinating residues, metal ions and substrates) was released and minimised for 15000 steps. The system was then heated from 0 to 303.15 K over 48500 steps, keeping a harmonic restraint of 10 $kcal.mol^{-1}.Å^{-2}$ on the backbone atoms and a restraint of 100 $kcal.mol^{-1}.Å^{-2}$ on the metal coordinating residues, metal ion and substrates, and using a Langevin thermostat. The system was then equilibrated in the NPT ensemble using the Nosé-Hoover Langevin piston pressure

control by decreasing slowly the restraints on the backbone to 0 and the restraints on the coordinating residues, metal ions and substrates to 10 kcal.mol^{-1}.Å^{-2} over 71500 steps and the system was further equilibrated for 1 ns keeping only the 10 kcal.mol^{-1}.Å^{-2} on the coordinating residues, metal ions and substrates relative positions as well as on their relative distances and angles. After equilibration, the resulting structure was minimised for 15000 steps keeping the same restraints.

The N-terminal tails (residues 1 to 22) of the resulting minimised structure were removed, a water sphere of 25 Å centred on the active site of chain B as well as a solvent shell of 3.5 Å around the protein were kept and all other water molecules removed. Ions that were not required to neutralise all the charges of the system were also removed.

6.4.3 QM/MM simulations

The resulting system was then used as a starting point for the QM/MM calculations using ChemShell. In order to study the influence of the protonation state of Cys102 which coordinates cobalt, two versions of the system were prepared, one in which the cysteine is protonated, and one in which it is not (CHARMM parameters and topology of unprotonated cysteine were obtained from Foloppe *et al.*[319]). The QM region consists in PEP, E4P, Co^{2+}, Lys 112 (link atom between C_γ and C_δ), which is involved in the proton transfer, the water molecule binding both PEP and E4P as well as the residues coordinating the metal ion: Cys102 (link atom between Cα and Cβ), His 282 (link atom between C_β and C_γ), Glu 316 (link atom between C_β and C_γ) and Asp 342 (link atom between C_α and C_β) of chain B.

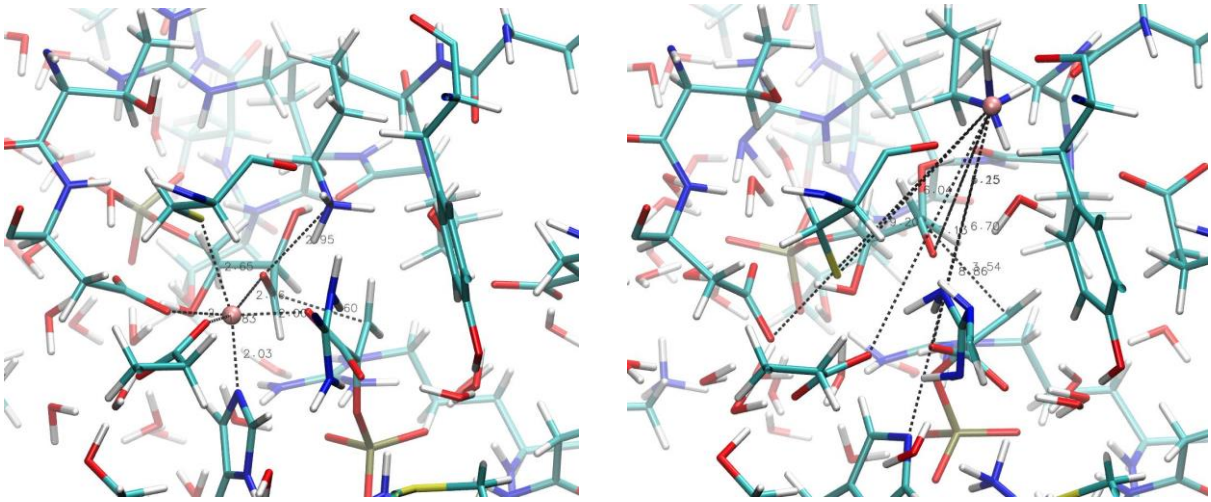


Figure 6.16. Geometry optimisation at the BP86/CHARMM22 level of theory (unprotonated cysteine). Left: starting geometry. Right: geometry after 230 steps (not converged).

Both unprotonated and protonated cysteine systems were optimised for about 230 steps with ChemShell using the BP86 functional and the small basis set def2-SV(P), the RI-J approximation with the DeMon/J Coulomb fitting auxiliary basis set. In both cases, the metal ion escaped from its coordination site, as observed with *Tma*DAH7P (see for example Figure 6.16).

To assess the effect of both the functional and the basis set on the outcome of the optimisation, additional tests were performed using the larger basis set def2-TZVP in conjunction with either the TPPSh functional which has recently been used successfully in the study of a cobalt-containing enzyme by QM/MM [320] or the PW6B95 which has been recommended for transition metal studies among others [321]. With these functionals and this basis set, calculations took longer, but eventually the escape of the metal ion was observed again (Figure 6.17).

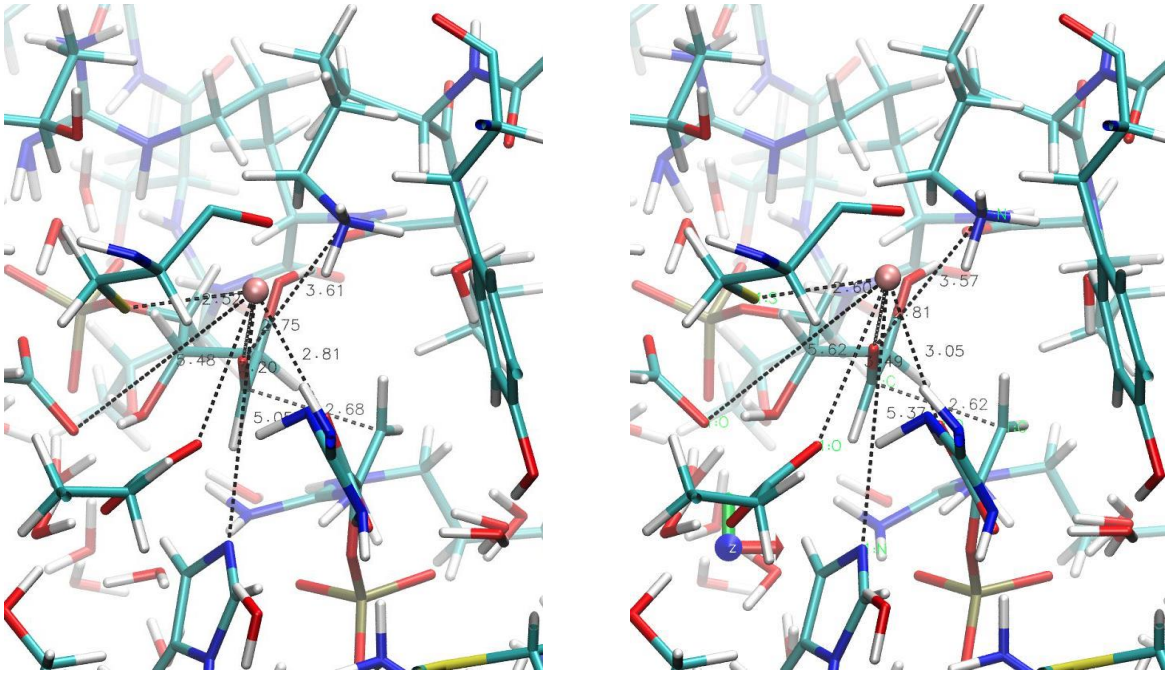


Figure 6.17. Geometry optimisations at various level of theory. Left: Geometry optimisation at the TPPSh/CHARMM22 level of theory after 89 steps (not converged). Right: Geometry optimisation at the PW6B95/CHARMM22 level of theory after 119 steps (not converged).

To see if the metal escaping is due to the MM protocol (minimisation, heating and equilibration) prior to the QM/MM calculations, geometry optimisation was also started directly from the crystal structure coordinates, removing G3P to only leave the coordinating water molecule. PEP was left in the MM region. Even under these conditions, the metal ion was found to leave its coordination site.

To check if the spin state of cobalt can have an impact on the result of the geometry optimization, Co^{2+} (modelled in its high spin state, multiplicity = 4, in the above tests) was also modelled in its low spin state (multiplicity = 2). Interestingly, with this multiplicity the cobalt ion remained in position during the optimisation. This is surprising as the pure QM geometry optimisation reported above were performed at a multiplicity of 4 using the same minimization algorithm and did not lead to any escape of the metal ion.

So far the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimisation method was used for the geometry optimisation. In order to see if this method could be the source of the problem, the simpler conjugate gradient (CG) approach was tested on the MD snapshot. Although the energy plateaus and does not reach convergence after 500 steps, the metal is still coordinated and overall, with this method, residues do not move as much as with the BFGS approach. However, during the CG minimization, the aspartate coordinating the metal ion was drifted away from Co^{2+}. Co^{2+} in turn get closer to the carboxyl group of PEP. This is observed for both the low and high spin states.

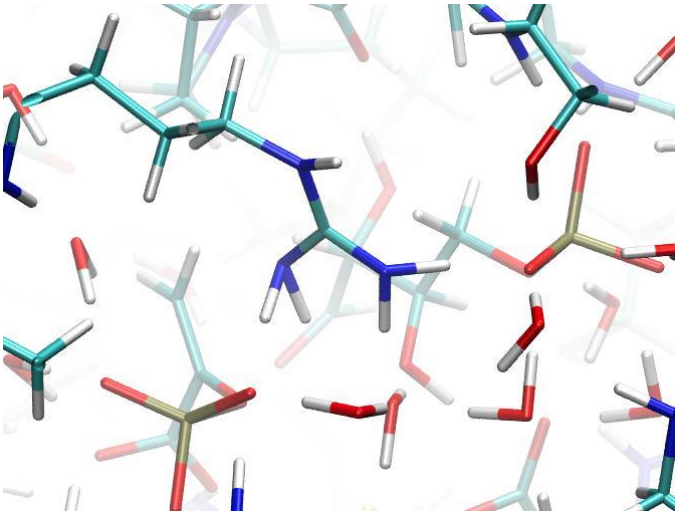


Figure 6.18. An example of deformation experienced by the guanidinium functional group of Arg residues during the BFGS optimisations.

To reach convergence after the CG steps, the BFGS algorithm was used to pursue the optimisation, however, this led to significant atomic displacements and unrealistic rearrangements. Moreover, during all the BFGS optimisations performed, major deformation of the Arg residues in the MM region that bind the phosphate groups can be observed, regardless of the inclusion of the phosphates in the QM region. These deformations concern

the hydrogens of the guanidinium that leave their planar geometry to act as if the lone pair on the nitrogen was localised instead of being conjugated with the double bond (Figure 6.18). The deformations are not observed during MM minimization with NAMD or during the minimization using the CG algorithm.

6.5 Discussion and perspectives

At this stage, because two starting structures (*Tma*DAH7PS, *Sce*DAH7PS) which present significant differences on how E4P is modelled and on the initial geometry of the coordinating residues were used, it is reasonable to state that the metal ion escape is not linked to the crystal structure or to the relative position of E4P, PEP, metal ion and coordinating residues. As all the key distances and angles were restrained, this is not due either to major deviations in the position of the metal ion, coordinating residues and substrates occurring during the MM simulation phase. The escape is also not affected by the nature of the metal ion (Co^{2+}, Mn^{2+} and Zn^{2+} tested) nor by the type or size of the basis set (3-21+G*, 6-31+G*, def2-SV(P) and def2-TZVP tested) or the type of DFT functional (BP86, B3LYP, TPPSh and PW6B95 tested).

Simulation of the metal centre in ORCA, with or without an implicit solvation model, does not lead to a metal ion escape. This could suggest that the escape observed in the QM/MM calculations is due to the effect of the residues neighbouring the QM region. However, the inclusion of a few additional residues for the pure QM calculations did not lead to the release of the metal ion. For completeness, it might be interesting to run the QM/MM optimisation including the residues that form a "second shell" around the metal centre. It may also make sense to include all the residues interacting with the substrates and see if a different fate for the

metal ion is observed. However, the inclusion of so many atoms will rapidly yield prohibitive calculation times.

Another possible reason for the metal escape might be that the coordinating residues are a too close to the metal due to crystal packing effects. Because throughout the MM minimization and equilibration, the distances between the metal ion and its coordinating residues are essentially kept at their crystal structure value, they might reflect more the constraint coming from the crystal packing than the more relaxed state expected in solution. However, removal of the restraints during the MM modelling leads to significant atomic deviations because of the partial inability of the force field to accurately model the interactions between the metal ion, coordinating residues and substrates. In order to have a more relaxed geometry prior to high-level QM/MM calculations, it might be interesting to model the active site using DFTB3 and run minimization and MD in CHARMM using this method. Furthermore, it can be checked if the escape of the metal is also observed with a method that relies on a different programme. In order to do that, conversion of Co^{2+} to Zn^{2+}, for which the parameters have been developed for DFTB3, is required.

Changing the multiplicity from 4 to 2 in the *Sce*DAH7PS calculations leads the metal to retain its initial position throughout the optimisation. If this is an encouraging result, it although raises a number of questions, especially if the spin state of the metal ion is so important, why pure QM calculations at a multiplicity of 4 did not lead to the escape of the metal ion.

Due to the aforementioned results, and since the escape of the metal ion was observed for the QM/MM calculations but not for the QM calculations alone, an error in the ChemShell programme cannot completely be ruled out. Moreover, the deformation observed for the

guanidinium functional group of Arg residues during the minimisation (Figure 6.18), and the large amplitude atomic displacements observed after only a few steps of BFGS optimisation also tends to point toward a problem in the DL-FIND algorithm which is used for the geometry optimisation in ChemShell. Therefore, the next steps would primarily entail discussing the guanidinium deformations as well as the escape of the metal ions when using the BFGS optimisation as implemented in DL-FIND with ChemShell developers and work with them on trying to solve this problem. Additionally, as noted previously, it might be interesting to try including more residues in the QM region and a test a QM/MM approach using the DFTB3 level of theory in CHARMM.

Unfortunately, due to the finite nature of a PhD project, this work could not be completed in the allowed time. Nonetheless, it paves the way to the successful modelling of the reaction catalysed by DAH7PS and provides a solid basis to further attempt to simulate the first step of the reaction by providing important clues about what is important in the modelling. Indeed, we have shown that PDB 1OF8 (*Sce*DAH7PS) represents, to date, the best starting point for the simulation by providing a better resolution of the active site than PDB 1RZM (*Tma*DAH7PS). Moreover, special attention should be given to the protonation state of the metal coordinating cysteine and that both the protonated and unprotonated forms should be tested. We have also highlighted the importance of the spin state of the metal ion as well as the inclusion of the metal coordinating residues into the QM region. Finally, the identification, via the QM/MM-MD simulations, of a water molecule interacting with both PEP and E4P that might help to orientate the carbonyl group into a reactive position, suggests this water should be included in the QM region for high-level calculations.

Furthermore, if the importance of this water molecule is confirmed, it provides an interesting starting point for the development of potential transition state analogues as active site inhibitors. Indeed, taken together, the water molecule and the transition state form a relatively rigid entity with the water molecule interacting with an arginine residue, the phosphate of PEP and occasionally with the hydroxyl of PEP (Figure 6.19). Therefore, mimicking this rigidity by designing a compound presenting one or more rings is likely to be an effective strategy if this water molecule is indeed important. The two phosphate groups would need to be replaced for improved cell penetration and stability, although finding phosphate isosteres is notoriously challenging [322]. Removal of the hydroxyl group from the E4P part of the analogue should improve the inhibition (as shown in [293]) as well as the bioavailability. Moreover, this region can be further rigidified by adding double bonds in order to mimic the correct position of the carbons due to the hydroxyl interacting with the protein in the transition state. However, a correct transition state should be modelled first, providing information on the optimal distances and angles required in the design of a transition state analogue.

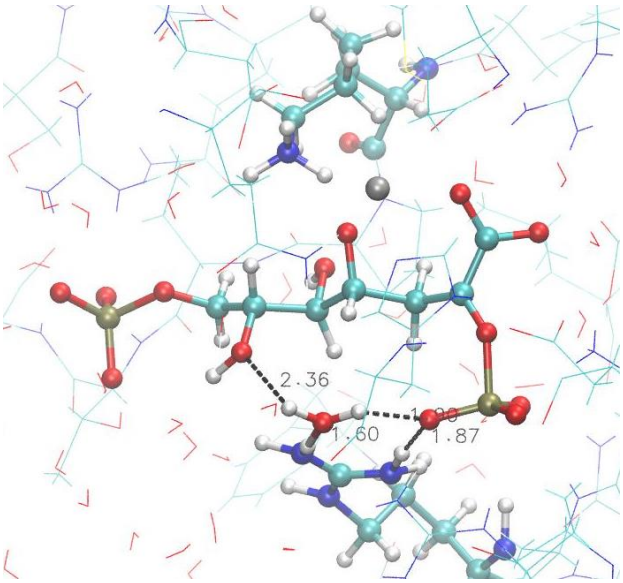


Figure 6.19. Structure of the transition state as obtained from the QM/MM-MD simulations with the water molecule.

Chapter 7 – Summary and perspectives on the different allosteric regulation mechanisms

Allostery and protein dynamics are intimately linked and traditionally three models are used to explain allosteric response following the binding of an allosteric inhibitor. *Induced fit*, for which binding of a ligand induces a conformational change in the protein toward an inactive conformation, *conformational selection* for which binding of a ligand stabilises high energy inactive conformations seldom sampled, and *dynamic allostery*, for which binding of a ligand modifies the frequency and amplitude of the protein motions leading to a reduced activity. Although it has been proposed that these mechanisms can coexist, e.g. conformational selection can be followed by induced fit, these models are usually described as three distinct, unrelated, regulation mechanisms.

The DAH7PS family, which catalyses the first committed step in the biosynthesis of aromatic amino acids, exhibits a unique variety of allosteric regulation mechanisms due to the great diversity of structural variations the members of this protein family exhibit. Phylogenetic analysis of the DAH7PS family (Chapter 2) suggests that allosteric regulation in DAH7PS originates from the acquisition of an N- or C-terminal regulatory domain obtained via gene fusion, essentially leading to the type Iβ DAH7PS subfamily, which also includes some

unregulated DAH7PS. The two other DAH7PS subfamilies, type Iα and type II, share the same catalytic barrel structure than type Iβ DAH7PS, despite relatively low sequence identity, especially in the case of type II. They are also equipped with an N-terminal allosteric decoration, but include in addition large loop insertions that have been proven to be critical for regulation. Phylogenetic analysis suggests that type Iα and type II DAH7PS might share a common ancestor, this common ancestor sharing, in turn, a common ancestor with type Iβ DAH7PS. It was found that there is a low evolutionary pressure on the N-terminal ACT domain that regulate certain type Iβ DAH7PS. Moreover, a clade of type Iβ enzymes closer to type Iα than any other type Iβ DAH7PS has been identified and possess loop extensions in the same regions than type Iα. This supports the idea of an ancestor of type Iα and type II DAH7PS which has evolved via mutations of a N-terminal regulatory domain and loop elongation in an early type Iβ enzyme.

When equipped with an ACT or CM N-terminal domain, allosteric regulation in type Iβ DAH7PS seems to occur via obstruction of the entrance to the catalytic site (Chapter 4 and Chapter 5). In the absence of allosteric ligand, the enzyme is in an active open form, whereas in the presence of an allosteric ligand the enzyme adopts an inactive closed form with the regulatory domains occluding the active site. The bulkiness of these regulatory domains lowers the catalytic activity of the enzyme when compared to a mutant lacking the regulatory domains. Molecular dynamics simulations and SAXS analysis have revealed that the active open form is able to sample conformations that resemble the inactive closed form, thereby suggesting a regulation mechanism based on conformational selection in which the inhibitor stabilises the inactive conformations. However, it is likely that the fully closed forms are, in fact, virtually inaccessible in the absence of ligand, which might suggest a partial induced fit mechanism

Indeed, in the case of an N-terminal ACT domain, one possibility is that sampling of the fully closed, inactive state, which is stable and presents a low relative free energy, may require passing high energy barriers and sampling high energy intermediate states that allow the transition from the open form conformations to the fully closed form. This accounts for the large degree of freedom of the regulatory domains and the associated variety of conformations they can sample, leading to a low probability to sample a fully closed state. The presence of Tyr may facilitate the interactions between adjacent ACT domains, lower the energy of the intermediate states, and therefore facilitating the complete closing of the regulatory domains. In fact, by stabilising the intermediate states that allow for the transition, Tyr may make the sampling of these states more frequent and thus the transition to the closed state more likely. Then opening of a back door via displacement of a key catalytic loop might destabilise the closed form, enabling the intermediate states to be sampled again, leading to the release of Tyr and the adoption of open conformations by the enzyme.

The situation may be similar in the case of an N-terminal CM domain, binding of chorismate or prephenate is likely to stabilise a conformation of the CM domains that presents extended helical regions in lieu of unstructured loops. This results in a more rigid intermediate state, which presents a lower energy in the presence of ligand since the entropically unfavourable folding of unstructured loops is counterbalanced by the protein ligand interactions which are enthalpically favourable. This intermediate state allows the CM domains to 'lean' on the DAH7PS catalytic barrels leading to the closed inactive form. Although the result of the MD simulations should provide more information on the exact nature of the mechanism, such conformational changes triggered by the binding of a ligand may correspond to the contribution of the induced fit mechanism to the global regulation mechanism. The presence of the ligand might not always be required to trigger the conformational changes and

in rare cases, the fully closed conformation might be sampled, however, as for the ACT domains, the sampling of intermediate states that are stabilised by the allosteric ligand increases the likelihood to reach the fully closed state.

Such an hybrid conformational selection / induced fit mechanism is illustrated in Figure 7.1. For an ensemble of conformations accessible to a protein, some conformations are only sampled very occasionally and briefly due to high energy barriers and high energy states (Figure 7.1. a). These conformations can serve, however, as intermediate states to reach other conformations. Since these high energy conformations are almost never sampled, they can be considered as virtually inaccessible. Binding of a ligand might make more stable some of the intermediate conformations by lowering their energy, leading to a population shift toward these intermediate conformations (Figure 7.1. b). A ligand would bind these intermediate conformations first. It is only once these intermediate states are liganded that a conformational change toward the previously virtually inaccessible states is possible (Figure 7.1c). In other words, binding of the allosteric ligand to the intermediate states enables a region of the conformational space that was previously almost inaccessible to be explored.

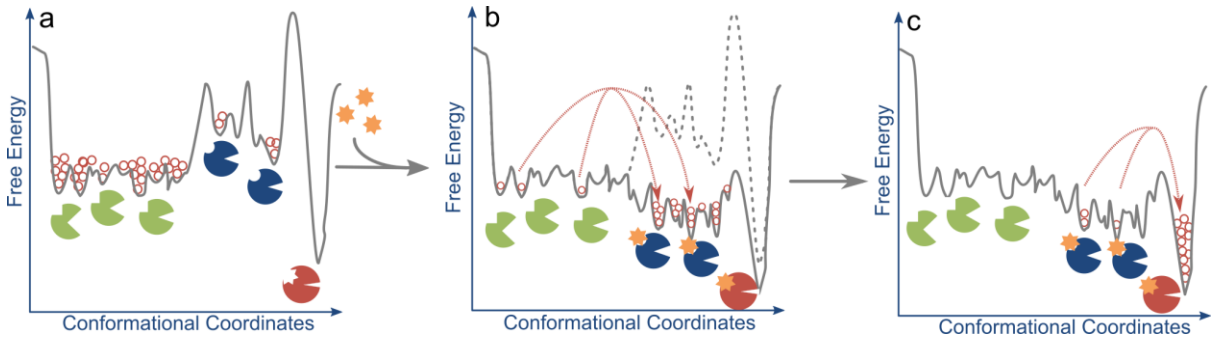


Figure 7.1. Schematic representation of the hybrid conformational selection / induced fit allosteric regulation mechanism possibly at play in N-terminal domain regulated type Iβ DAH7PS. Each bead represents the relative population of a given state. In the absence of allosteric ligand, most conformations correspond to the active open forms (green). A few conformations (blue) can serve as intermediates to give access

to the fully closed state (red), but these conformations are not often sampled, and the fully closed form is therefore virtually inaccessible (a). The addition of an allosteric ligand (orange stars) enables to stabilise the intermediate conformations by reducing their relative energy and the associated barriers (b). As a result, the intermediate states are sampled often and the fully closed state is more readily accessible and becomes the most populated conformation (c).

Analysis of allostery in type Iα DAH7PS (Chapter 3) reveals a regulation mechanism primarily based on dynamic allostery, with the catalytic activity in type Iα DAH7PS being drastically reduced by an increased flexibility of a key catalytic loop and changes in the relative position of the side chain of a number of catalytic residues. Analysis of the calculated pK_a variations demonstrated the importance of ionisable residues in the allosteric regulation. This analysis signifies that the regulation mechanism of type Iα DAH7PS relies on small, localised conformational changes of residue side chains that lead to the formation or cleavage of non-covalent interactions that are responsible for the change in dynamics. This confirms that the existence of dynamic allostery in the complete absence of conformational change is extremely unlikely. Moreover, the existence of small conformational changes would, therefore, suggests that an underlying conformational selection mechanism is also at play. In this case there might be a hybrid conformational selection / dynamic allostery mechanism.

As illustrated in Figure 7.2, dynamic allostery can indeed be explained in terms of conformational changes, as the fluctuations around an averaged conformation can simply be seen as the interconversion between an ensemble of closely related conformations, which all have very similar energies (Figure 7.2.a.). Binding of a ligand might stabilise a subset of these conformations by lowering their energy (Figure 7.2.b.), leading to a population shift toward these more stable states. Overall, the amplitude of the motions has been reduced, because the number of accessible conformations has been diminished (Figure 7.2.c.).

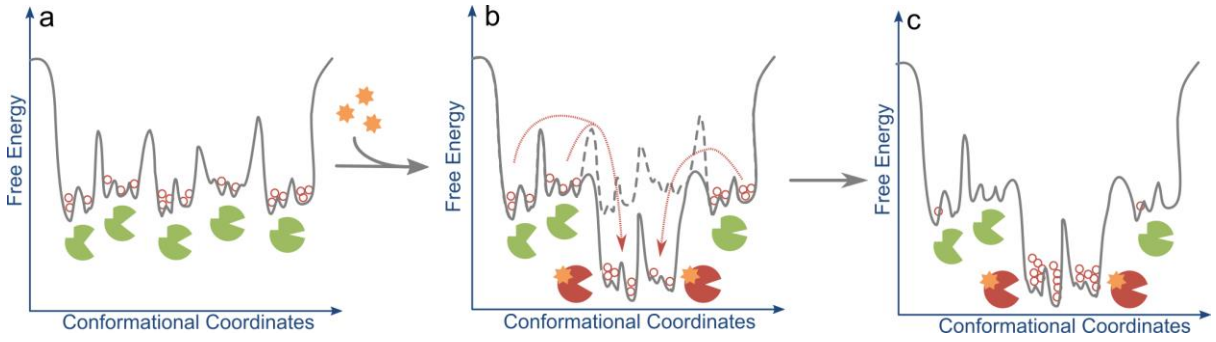


Figure 7.2. Schematic representation of dynamic allostery using conformational selection. Each bead represents the relative population of a given state. In the absence of ligand interconversion between closely related conformations (green) is facile and the enzyme is able to perform its reaction by being able, for example, to reach a broadly open conformation in order to accept the substrates and release the products and a mostly closed conformation that enables the reaction to occur (a). In the presence of a ligand (orange stars), a few of these conformations are stabilised (red), leading to a population shift (b). The enzyme is not able anymore to reach the broadly open and mostly closed conformations and cannot perform the catalytic reaction (c).

Overall, these results suggest that allosteric regulation mechanisms can proceed via a subtle combination of the three classical models of allostery, conformational selection, induced fit and dynamic allostery. Refinement of regulation mechanisms both in DAH7PS and in other systems will, therefore, help unravel some of the complexity and interrelatedness in the allosteric regulation mechanisms, helping to contribute to a robust, widely accepted universal model of allostery.

Appendix A – Supplementary movies

All the movies are deposited on Figshare and can be accessed by following this link:

https://figshare.com/s/823892477b055ad53161:

Movie 4.1. MD simulation of *Tma*DAH7PS Open_run1, side view. The trajectory corresponds to a MD simulation of 635 ns at 363 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red and $\beta_8\alpha_8$ loops are coloured in green.

Movie 4.2. MD simulation of *Tma*DAH7PS Open_run1, top view. The trajectory corresponds to a MD simulation of 635 ns at 363 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red and $\beta_8\alpha_8$ loops are coloured in green.

Movie 4.3. MD simulation of *Tma*DAH7PS Open_300K, side view. The trajectory corresponds to a MD simulation of 335 ns at 300 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red and $\beta_8\alpha_8$ loops are coloured in green.

Movie 4.4. MD simulation of *Tma*DAH7PS Open_300K, top view. The trajectory corresponds to a MD simulation of 335 ns at 300 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red and $\beta_8\alpha_8$ loops are coloured in green.

Movie 4.5. MD simulation of *Tma*DAH7PS Closed_run1, side view. The trajectory corresponds to a MD simulation of 335 ns at 363 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red, $\beta_8\alpha_8$ loops are coloured in green and allosteric Tyr are coloured in purple.

Movie 4.6. MD simulation of *Tma*DAH7PS Closed_run1, top view. The trajectory corresponds to a MD simulation of 335 ns at 363 K. Catalytic barrels of tetrameric *T. maritima* DAH7PS are shown in white, regulatory domains are shown in blue, $\beta_2\alpha_2$ loops are coloured in red, $\beta_8\alpha_8$ loops are coloured in green and allosteric Tyr are coloured in purple.

Movie 5.1. Morphing between the open form and closed form of *Gsp*DAH7PS. The CM regulatory domains are shown in green and the catalytic barrels in blue. Morphing was obtained with the Morph command of PyMOL over 60 states and using five sculpting refinement cycles. The movie includes both the open to closed and closed to open morphing.

Movie 6.1. QM/MM modelling of the bond formation between PEP and E4P in *Tma*DAH7PS at the AM1/d-PhoT/CHARMM-22 level of theory. Result of the umbrella sampling at the AM1/d-PhoT/CHARMM-22 level of theory with a QM region consisting of PEP (right), E4P (left) (without its phosphate group), Zn^{2+} (grey sphere), part of Lys131 (top) and the water molecule.

Movie 6.2. QM/MM energy minimization of *Tma*DAH7PS at the BP86/3-21+G*/d-PhoT/CHARMM-22 level of theory for the QM1 region. QM1 includes Mn^{2+} (pink sphere), PEP, E4P (without its phosphate) and Lys131 (cut between Cβ and Cγ). Cys102 is deprotonated.

Movie 6.3. QM/MM energy minimization of *Tma*DAH7PS at the BP86/3-21+G*/d-PhoT/CHARMM-22 level of theory for the QM2 region. QM2 includes QM1 + water molecule interacting with both a hydroxyl of E4P and the phosphate of PEP identified as being important in the QM/MM-MD simulations. Cys102 is deprotonated.

Movie 6.4. QM/MM energy minimization of *Tma*DAH7PS at the BP86/3-21+G*/d-PhoT/CHARMM-22 level of theory for the QM3 region. QM3 includes QM2 + Mn^{2+} coordinating residues (Cys102, His272, Glu298 and Asp309). Cys102 is deprotonated.

Movie 6.5. QM/MM energy minimization of *Tma*DAH7PS at the BP86/3-21+G*/d-PhoT/CHARMM-22 level of theory for the QM3 region with Cys102 protonated. QM3 includes QM2 + Mn^{2+} coordinating residues (Cys102, His272, Glu298 and Asp309). Cys102 is deprotonated.

Appendix B – Supporting information for Chapter 3

Table S2. Calculated Phe dissociation constants from ITC experiments

	Site 1	Site 2	Model
pH 7	**3.3 ± 0.8 µM**	**11 ± 1.3 µM**	**Sequential two site**
pH7.8	**2.0 ± 0.7 µM**	**8.6 ± 1.2 µM**	**Sequential two site**

Table S2. Data Collection and refinement

	***Nme*DAH7PS**
	L-Phenylalanine
Data Collection	
Crystal system; space group	Monoclinic, $P12_11$
Unit cell parameters	
***a, b, c* (Å)**	73.61, 143.50, 75.18
α, β, γ (°)	90, 96.18, 90
Resolution range (Å)	47.82-2.19 (2.23-2.19)
Measurements	301997
Unique reflections	79522
Redundancy	3.8
Completeness (%)	99.9 (100.0)
***I*/σ (*I*)**	9.3 (1.2)
R_{merge}	0.084
$CC_{1/2}$	0.50
Wilson *B* value (Å)^2	39.81
Matthews coefficient	2.60
Refinement	
R_{work}	0.1890
R_{free}	0.2357
Chain length	351
Observed number of residues	345 (All Chains)
Water molecules	330
Other (Mn^{2+}, PEG, SO_4^-)	9

Ligand	4
Mean *B* (Å)^2	
Protein	48.59
Water	43.82
Other	67.46
Phe	37.13
R.m.s.d from target values	
Bond lengths (Å)	0.0095
Bond angles	1.2712
Dihedral angles	0.0690
Ramachandran	
Preferred (%)	97.69
Allowed (%)	1.73
Outliers (%)	0.58
PDB Entry	4UC5

Table S3. Substrate affinities at pH 6.8 and 8.3

	pH 6.8	**pH 8.3**
K_m^{E4P} (µM)	**36.7 ± 1.8**	**106 ± 15**
K_m^{PEP} (µM)	**15.4 ± 0.9**	**7.0 ± 0.6**

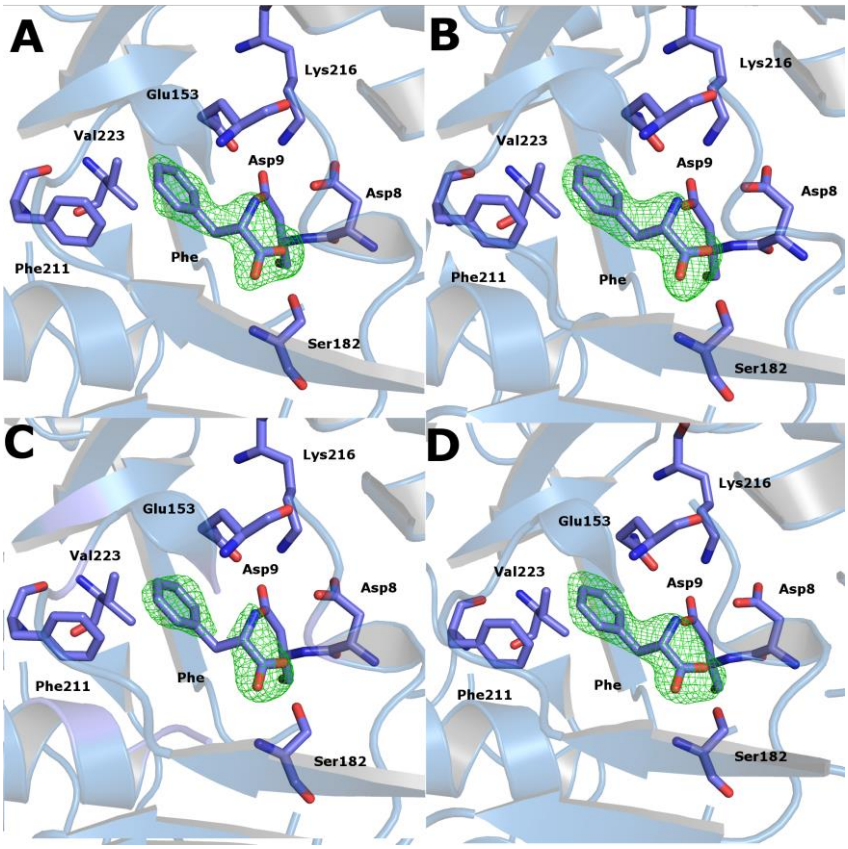


Figure S1. $|F_o\text{-}F_c|$ maps for L-phenylalanine bound at the allosteric site contoured at 3σ. The four chains of *Nme*DAH7PS are indicated. Top left: chain A, top right; chain B, bottom left: chain C and bottom right: chain D.

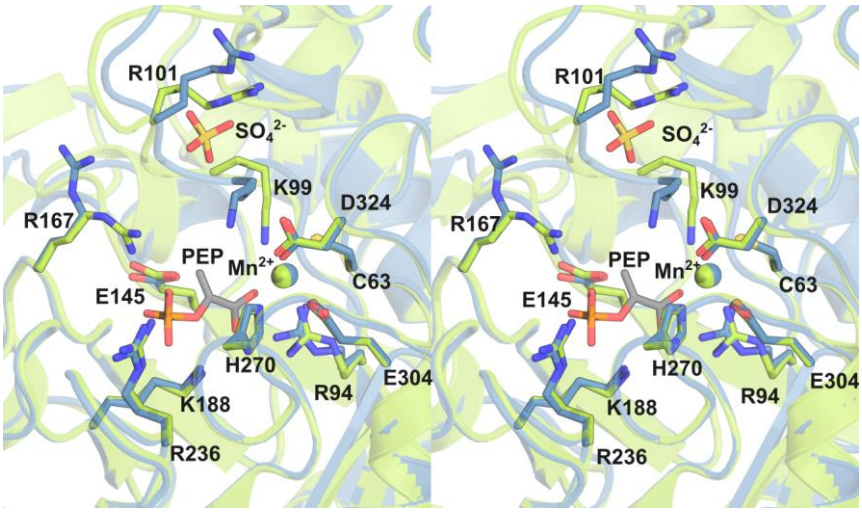


Figure S2. Stereoview of the overlay of the active site of *Nme*DAH7PS for the apo (green) and Phe-bound (blue) forms. Catalytically important residues are shown with sticks. Manganese ions are represented with a sphere.

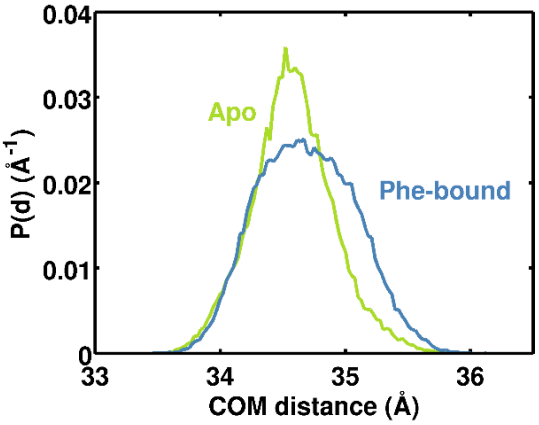


Figure S3. Distribution of the distance between the center of mass (COM) of each monomer for the apo (green) and Phe-bound (blue) MD simulations.

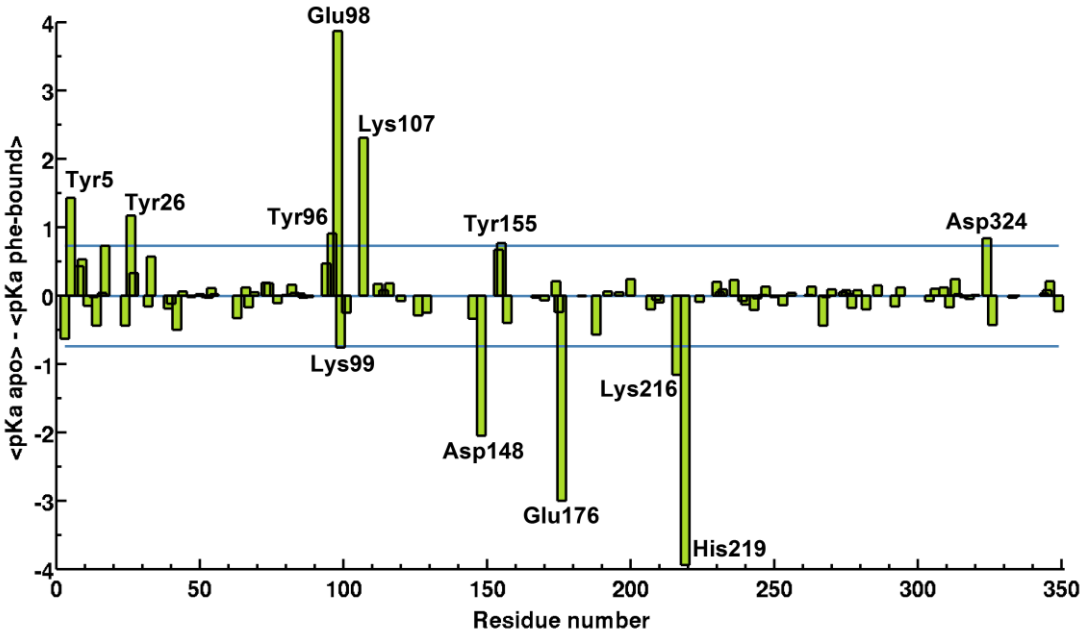


Figure S4. Difference between the average pK_a difference of ionizable residues between apo and Phe-bound forms of *Nme*DAH7PS, over the course of the MD simulations. The middle blue line represents the average difference in pK_a between the apo and Phe-bound form and the top and bottom blue lines correspond to the average ± one standard deviation. Residues that present a major pK_a shift and therefore might be involved in the allosteric communication pathway have been selected on the basis of their distance from the average ± one standard deviation.

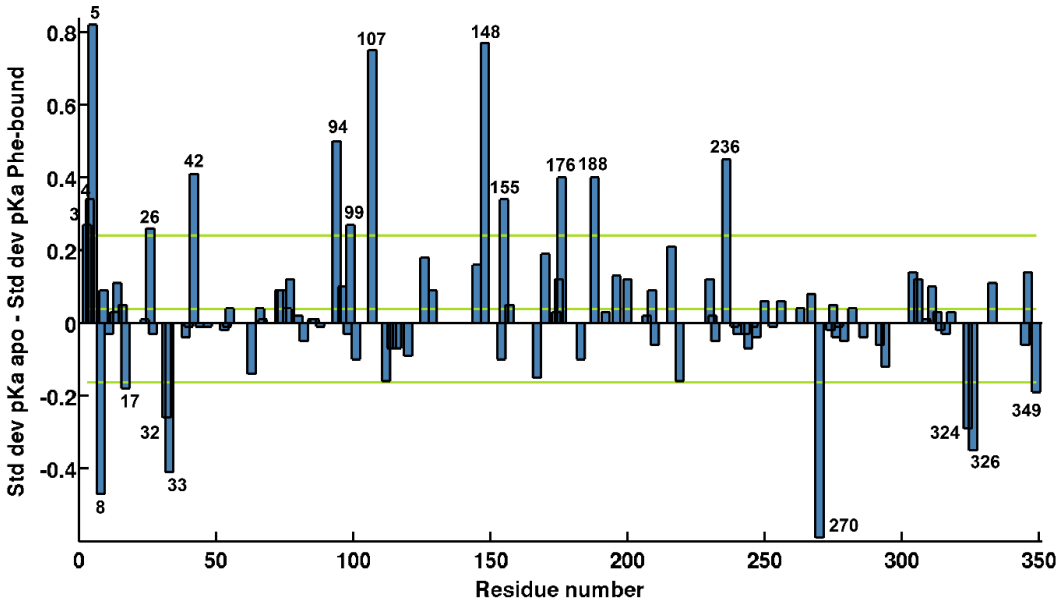


Figure S5. Difference between the standard deviation of the pK_a of the ionizable residues of the apo and Phe-bound forms, over the course of the MD simulations. The middle green line represents the average difference and the upper and lower green lines correspond to the average ± one standard deviation. Residues that might be involved in the allosteric communication pathway have been selected on the basis of their distance from the average ± one standard deviation.

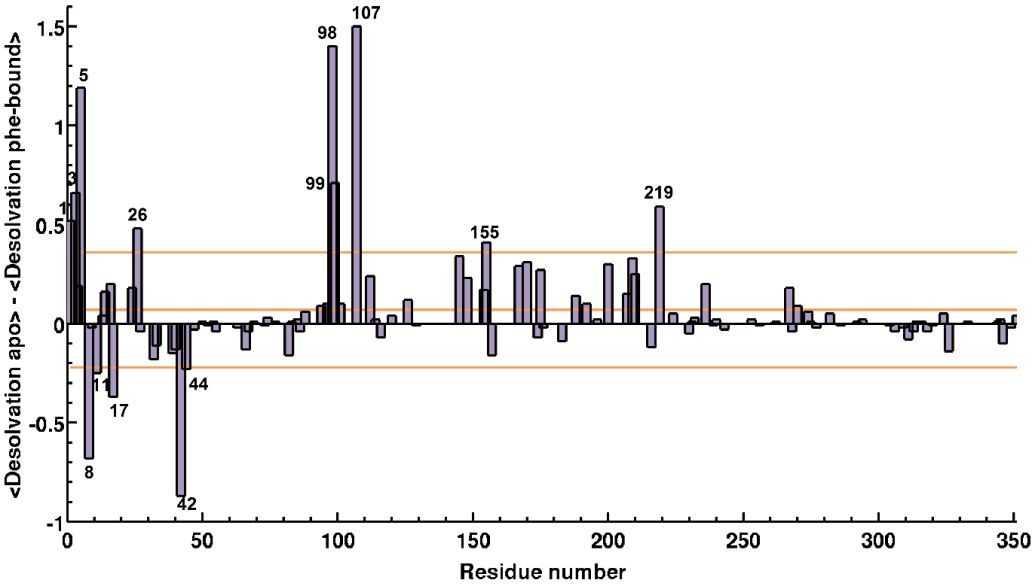


Figure S6. Difference between the desolvation values of the ionizable residues of the apo and Phe-bound forms, averaged over the course of the MD simulations. The desolvation difference is given in log units. The larger the positive difference, the more buried the residue is in the Phe-bound structure. The middle orange line represents the average difference and the two other orange lines correspond to the average ± one standard deviation.

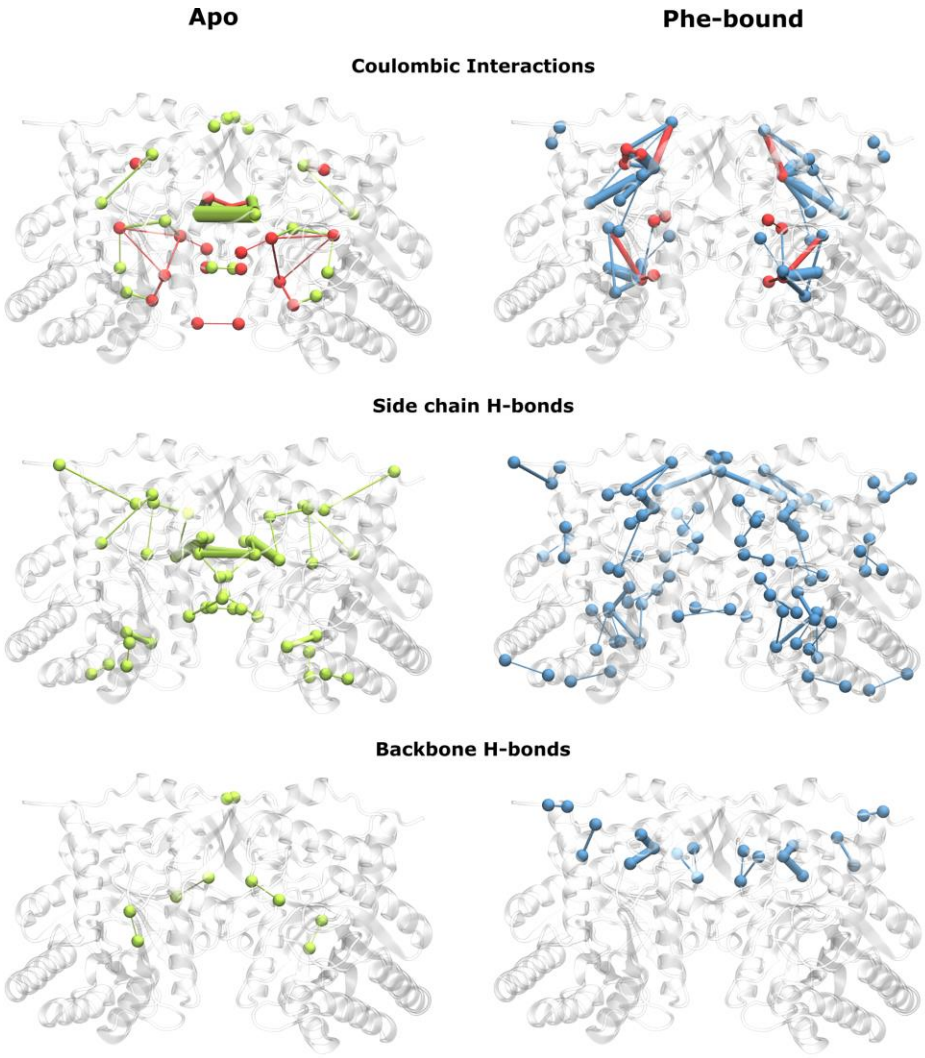


Figure S7. Detailed differential maps of interactions between the apo and Phe-bound forms. Details of the coulombic interactions, side chain H-bounds (between the ionizable group of a residue and the hydrogen donor or acceptor function present on the side chain of another residue) and backbone H-bonds (between the ionizable group of a residue and the backbone of another) that are only present or stronger in the apo (left column) or the Phe-bound (right columns) trajectories. Nodes represent interacting residues and the edges between nodes indicate the interactions, with the diameter of the edge accounting for the strength of the interaction. For the apo form, favorable interactions are in blue and unfavorable ones are in orange, and for the Phe-bound form, favorable interactions are in red and unfavorable ones are in purple.

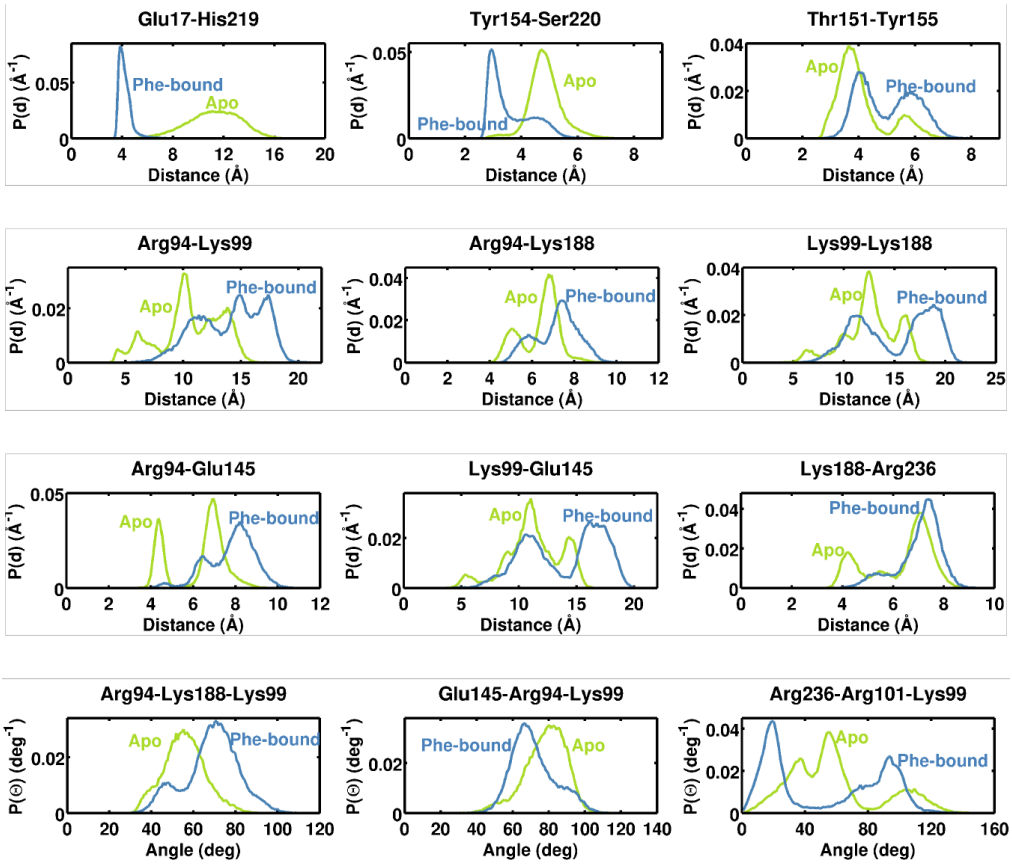


Figure S8. Distribution of the distances between ionizable groups of interest. Frequency distribution of the distances for the apo (green) and Phe-bound (blue) forms. The distances are calculated using the hydroxyl oxygen of Ser, Tyr, Thr the center of the two carboxyl oxygens of Glu and Asp, the nitrogen of the ε-amino group of Lys, the center of the guanidinium ion of Arg and the center of the two nitrogens of the imidazole ring of His.

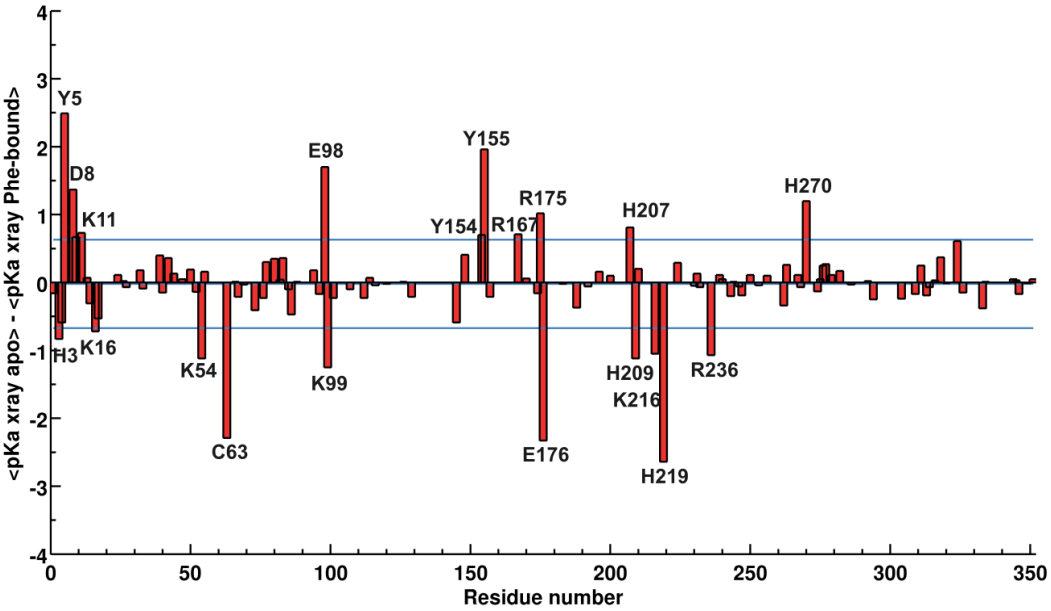


Figure S9. Difference between the pK_a difference of ionizable residues between apo and Phe-bound models generated by reconstructing missing residues and side chains from the apo and Phe-bound X-ray crystal structures

of *Nme*DAH7PS. The pK_a values were obtained by averaging by model and by chain. The middle blue line represents the average difference in pK_a between the apo and Phe-bound form and the top and bottom blue lines correspond to the average ± one standard deviation.

References

1. Weber G: **Energetics of ligand binding to proteins**. *Advances in Protein Chemistry* 1975, **29**:1-83.

2. Kern D, Zuiderweg ER: **The role of dynamics in allosteric regulation**. *Current Opinion in Structural Biology* 2003, **13**:748-757.

3. Roy R, Hohng S, Ha T: **A practical guide to single-molecule FRET**. *Nature Methods* 2008, **5**:507-516.

4. Kleckner IR, Foster MP: **An introduction to NMR-based approaches for measuring protein dynamics**. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2011, **1814**:942-968.

5. Engen JR: **Analysis of Protein Conformation and Dynamics by Hydrogen/Deuterium Exchange MS**. *Analytical Chemistry* 2009, **81**:7870-7875.

6. Martinez-Julvez M, Abian O, Vega S, Medina M, Velazquez-Campoy A: **Studying the Allosteric Energy Cycle by Isothermal Titration Calorimetry**. In *Allostery: Methods and Protocols*. Edited by Fenton WA: Springer New York; 2012:53-70. 10.1007/978-1-61779-334-9_4

7. Schellman JA, Schellman CG: **Kaj Ulrik Linderstrom-Lang (1896-1959)**. *Protein Science* 1997, **6**:1092-1100.

8. Linderstrøm-Lang KUS, J. A.: **Protein Structure and Enzyme Activity**. In *The Enzymes*, edn 2nd. Edited by Myrbäk BL: Academic Press; 1959:443-510 vol 1.]

9. Lakowicz JR, Weber G: **Quenching of protein fluorescence by oxygen. Detection of structural fluctuations in proteins on the nanosecond time scale**. *Biochemistry* 1973, **12**:4171-4179.

10. Grinvald A, Steinberg IZ: **Fast relaxation processes in a protein revealed by the decay kinetics of tryptophan fluorescence**. *Biochemistry* 1974, **13**:5170-5178.

11. Eftink MR, Ghiron CA: **Dynamics of a protein matrix revealed by fluorescence quenching**. *Proceedings of the National Academy of Sciences* 1975, **72**:3290-3294.

12. Englander SW, Mauel C: **Hydrogen Exchange Studies of Respiratory Proteins: II.Detection of Discrete, Ligand-induced Changes in Hemoglobin**. *Journal of Biological Chemistry* 1972, **247**:2387-2394.

13. Saviotti ML, Galley WC: **Room Temperature Phosphorescence and the Dynamic Aspects of Protein Structure**. *Proceedings of the National Academy of Sciences* 1974, **71**:4154-4158.

14. Oster O, Neireiter GW, Clouse AO, Gurd FR: **Deuterium nuclear magnetic resonance of deuterium-labeled diacetyldeuterohemin incorporated into sperm whale myoglobin**. *Journal of Biological Chemistry* 1975, **250**:7990-7996.

15. McCammon JA, Gelin BR, Karplus M: **Dynamics of folded proteins**. *Nature* 1977, **267**:585-590.

16. Gurd FRN, Rothges TM: **Motions in Proteins**. In *Advances in Protein Chemistry*. Edited by C.B. Anfinsen JTE, Frederic MR: Academic Press; 1979:73-165. vol Volume 33.]

17. Careri G, Fasella P, Gratton E, Jencks WP: **Statistical Time Events in Enzymes: A Physical Assessmen**. Edited by; 1975. 10.3109/10409237509102555

18. G Careri, P Fasella a, Gratton E: **Enzyme Dynamics: The Statistical Physics Approach**. *Annual Review of Biophysics and Bioengineering* 1979, **8**:69-97.

19. Frauenfelder H, Sligar SG, Wolynes PG: **The energy landscapes and motions of proteins**. *Science* 1991, **254**:1598-1603.

20. Henzler-Wildman K, Kern D: **Dynamic personalities of proteins**. *Nature* 2007, **450**:964-972.

21. Brooks CL, Onuchic JN, Wales DJ: **Taking a Walk on a Landscape**. *Science* 2001, **293**:612-613.

22. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR: **Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations**. *Proceedings of the National Academy of Sciences* 2009, **106**:19011-19016.

23. Ma B, Kumar S, Tsai C-J, Nussinov R: **Folding funnels and binding mechanisms**. *Protein Engineering* 1999, **12**:713-720.

24. Tsai C-J, Kumar S, Ma B, Nussinov R: **Folding funnels, binding funnels, and protein function**. *Protein Science* 1999, **8**:1181-1190.

25. Tsai C-J, Ma B, Nussinov R: **Folding and binding cascades: Shifts in energy landscapes**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**:9970-9972.

26. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D: **A hierarchy of timescales in protein dynamics is linked to enzyme catalysis**. *Nature* 2007, **450**:913-916.

27. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R: **Folding and binding cascades: dynamic landscapes and population shifts**. *Protein Science : A Publication of the Protein Society* 2000, **9**:10-19.

28. Motlagh HN, Wrabl JO, Li J, Hilser VJ: **The ensemble nature of allostery**. *Nature* 2014, **508**:331-339.

29. del Sol A, Tsai CJ, Ma B, Nussinov R: **The origin of allosteric functional modulation: multiple pre-existing pathways**. *Structure* 2009, **17**:1042-1050.

30. Nussinov R, Tsai C-J, Ma B: **The Underappreciated Role of Allostery in the Cellular Network**. *Annu Rev Biophys* 2013, **42**:169-189.

31. Traut T: *Allosteric regulatory enzymes*10.1007/978-0-387-72891-9: Springer Science 2008.

32. Nussinov R, Tsai CJ: **Allostery in disease and in drug discovery**. *Cell* 2013, **153**:293-305.

33. Fischer E: **Einfluss der Configuration auf die Wirkung der Enzyme**. *Berichte der deutschen chemischen Gesellschaft* 1894, **27**:2985-2993.

34. Koshland DE: **Application of a Theory of Enzyme Specificity to Protein Synthesis**. *Proceedings of the National Academy of Sciences* 1958, **44**:98-104.

35. Koshland DE, Jr., Nemethy G, Filmer D: **Comparison of experimental binding data and theoretical models in proteins containing subunits**. *Biochemistry* 1966, **5**:365-385.

36. Monod J, Wyman J, Changeux JP: **On the Nature of Allosteric Transitions: A Plausible Model**. *Journal of Molecular Biology* 1965, **12**:88-118.

37. Rubin MM, Changeux J-P: **On the nature of allosteric transitions: Implications of non-exclusive ligand binding**. *Journal of Molecular Biology* 1966, **21**:265-274.

38. Cui Q, Karplus M: **Allostery and cooperativity revisited**. *Protein Science* 2008, **17**:1295-1307.

39. Cooper A, Dryden DTF: **Allostery without conformational change**. *European Biophysics Journal* 1984, **11**:103-109.

40. Goodey NM, Benkovic SJ: **Allosteric regulation and catalysis emerge via a common route**. *Nature Chemical Biology* 2008, **4**:474-482.

41. Boehr DD, Nussinov R, Wright PE: **The role of dynamic conformational ensembles in biomolecular recognition**. *Nature Chemical Biology* 2009, **5**:789-796.

42. Gunasekaran K, Ma B, Nussinov R: **Is allostery an intrinsic property of all dynamic proteins?** *Proteins: Structure, Function, and Bioinformatics* 2004, **57**:433-443.

43. Jusuf S, Loll PJ, Axelsen PH: **Configurational Entropy and Cooperativity between Ligand Binding and Dimerization in Glycopeptide Antibiotics**. *Journal of the American Chemical Society* 2003, **125**:3988-3994.

44. Popovych N, Sun S, Ebright RH, Kalodimos CG: **Dynamically driven protein allostery**. *Nature Structural & Molecular Biology* 2006, **13**:831-838.

45. Tsai C-J, del Sol A, Nussinov R: **Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play**. *Journal of Molecular Biology* 2008, **378**:1-11.

46. Smock RG, Gierasch LM: **Sending signals dynamically**. *Science* 2009, **324**:198-203.

47. Changeux JP, Edelstein S: **Conformational selection or induced fit? 50 years of debate resolved**. *F1000 Biology Reports* 2011, **3**:19-19.

48. Formaneck MS, Ma L, Cui Q: **Reconciling the "old" and "new" views of protein allostery: A molecular simulation study of chemotaxis Y protein (CheY)**. *Proteins: Structure, Function, and Bioinformatics* 2006, **63**:846-867.

49. Vogt AD, Di Cera E: **Conformational Selection or Induced Fit? A Critical Appraisal of the Kinetic Mechanism**. *Biochemistry* 2012, **51**:5894-5902.

50. Vogt AD, Di Cera E: **Conformational Selection Is a Dominant Mechanism of Ligand Binding**. *Biochemistry* 2013, **52**:5723-5729.

51. Vogt AD, Pozzi N, Chen Z, Di Cera E: **Essential role of conformational selection in ligand binding**. *Biophysical Chemistry* 2014, **186**:13-21.

52. Weikl TR, Paul F: **Conformational selection in protein binding and function**. *Protein Science* 2014, **23**:1508-1518.

53. Vega S, Abian O, Velazquez-Campoy A: **On the link between conformational changes, ligand binding and heat capacity**. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2015, 10.1016/j.bbagen.2015.10.010.

54. Zhou H-X: **From Induced Fit to Conformational Selection: A Continuum of Binding Mechanism Controlled by the Timescale of Conformational Transitions**. *Biophysical Journal* 2010, **98**:L15-L17.

55. Greives N, Zhou H-X: **Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit**. *Proceedings of the National Academy of Sciences* 2014, **111**:10197-10202.

56. Bosshard HR: **Molecular Recognition by Induced Fit: How Fit is the Concept?** *Physiology* 2001, **16**:171-173.

57. Wlodarski T, Zagrovic B: **Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin**. *Proceedings of the National Academy of Sciences* 2009, **106**:19346-19351.

58. Nussinov R, Tsai C-J: **Allostery without a conformational change? Revisiting the paradigm**. *Current Opinion in Structural Biology* 2015, **30**:17-24.

59. Zhuravleva A, Gierasch LM: **Substrate-binding domain conformational dynamics mediate Hsp70 allostery**. *Proceedings of the National Academy of Sciences* 2015, 10.1073/pnas.1506692112.

60. Wrabl JO, Gu J, Liu T, Schrank TP, Whitten ST, Hilser VJ: **The role of protein conformational fluctuations in allostery, function, and evolution**. *Biophysical Chemistry* 2011, **159**:129-141.

61. Hilser VJ, Wrabl JO, Motlagh HN: **Structural and Energetic Basis of Allostery**. *Annual Review of Biophysics* 2012, **41**:585-609.

62. Ansari A, Berendzen J, Bowne SF, Frauenfelder H, Iben IE, Sauke TB, Shyamsunder E, Young RD: **Protein states and proteinquakes**. *Proceedings of the National Academy of Sciences of the United States of America* 1985, **82**:5000-5004.

63. Levantino M, Schirò G, Lemke HT, Cottone G, Glownia JM, Zhu D, Chollet M, Ihee H, Cupane A, Cammarata M: **Ultrafast myoglobin structural dynamics observed with an X-ray free-electron laser**. *Nat Commun* 2015, **6**.

64. Barends TR, Foucar L, Ardevol A, Nass K, Aquila A, Botha S, Doak RB, Falahati K, Hartmann E, Hilpert M, et al.: **Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation**. *Science* 2015, **350**:445-450.

65. Miyashita O, Onuchic JN, Wolynes PG: **Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins**. *Proceedings of the National Academy of Sciences* 2003, **100**:12570-12575.

66. Leitner DM: **Energy Flow in Proteins**. *Annual Review of Physical Chemistry* 2008, **59**:233-259.

67. Li G, Magana D, Dyer RB: **Anisotropic energy flow and allosteric ligand binding in albumin**. *Nat Commun* 2014, **5**.

68. Di Paola L, Giuliani A: **Protein contact network topology: a natural language for allostery**. *Current Opinion in Structural Biology* 2015, **31**:43-48.

69. Royer WE, Pardanani A, Gibson QH, Peterson ES, Friedman JM: **Ordered water molecules as key allosteric mediators in a cooperative dimeric hemoglobin**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**:14526-14531.

70. Feher VA, Durrant JD, Van Wart AT, Amaro RE: **Computational approaches to mapping allosteric pathways**. *Current Opinion in Structural Biology* 2014, **25**:98-103.

71. Tsai CJ, Del Sol A, Nussinov R: **Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms**. *Molecular Biosystems* 2009, **5**:207-216.

72. Tsai C-J, Nussinov R: **A Unified View of "How Allostery Works"**. *PLoS Computational Biology* 2014, **10**:e1003394.

73. Hilser VJ: **An Ensemble View of Allostery**. *Science* 2010, **327**:653-654.

74. Teilum K, Olsen J, Kragelund B: **Functional aspects of protein flexibility**. *Cellular and Molecular Life Sciences* 2009, **66**:2231-2247.

75. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE: **Biomolecular simulation: a computational microscope for molecular biology**. *Annu Rev Biophys* 2012, **41**:429-452.

76. Jensen F: *Introduction to Computational Chemistry*: John Wiley & Sons; 2006.

77. Adcock SA, McCammon JA: **Molecular dynamics: survey of methods for simulating the activity of proteins**. *Chemical Reviews* 2006, **106**:1589-1615.

78. Leach A: *Molecular Modelling: Principles and Applications* edn 2nd: Prentice Hall; 2001.

79. Frenkel D, Smit B: *Understanding Molecular Simulation*: Academic Press, Inc.; 2001.

80. Durrant J, McCammon JA: **Molecular dynamics simulations and drug discovery**. *BMC Biology* 2011, **9**:71-71.

81. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA: **Development and testing of a general amber force field**. *Journal of Computational Chemistry* 2004, **25**:1157-1174.

82. MacKerell AD, Jr., Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, et al.: **All-atom empirical potential for molecular modeling and dynamics studies of proteins**. *Journal of Physical Chemistry B* 1998, **102**:3586-3616.

83. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF: **A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6**. *Journal of Computational Chemistry* 2004, **25**:1656-1676.

84. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ: **The Amber biomolecular simulation programs**. *Journal of Computational Chemistry* 2005, **26**:1668-1688.

85. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: **CHARMM: A program for macromolecular energy, minimization, and dynamics calculations**. *Journal of Computational Chemistry* 1983, **4**:187-217.

86. Hess B, Kutzner C, van der Spoel D, Lindahl E: **GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation**. *Journal of Chemical Theory and Computation* 2008, **4**:435-447.

87. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: **GROMACS: Fast, flexible, and free**. *Journal of Computational Chemistry* 2005, **26**:1701-1718.

88. Christen M, Hünenberger PH, Bakowies D, Baron R, Bürgi R, Geerke DP, Heinz TN, Kastenholz MA, Kräutler V, Oostenbrink C, et al.: **The GROMOS software for biomolecular simulation: GROMOS05**. *Journal of Computational Chemistry* 2005, **26**:1719-1751.

89. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K: **Scalable molecular dynamics with NAMD**. *Journal of Computational Chemistry* 2005, **26**:1781-1802.

90. Ovchinnikov V, Karplus M: **Analysis and Elimination of a Bias in Targeted Molecular Dynamics Simulations of Conformational Transitions: Application to Calmodulin**. *The Journal of Physical Chemistry B* 2012, **116**:8584-8603.

91. van der Vaart A, Karplus M: **Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations**. *Journal of Chemical Physics* 2007, **126**:164106-164106.

92. Hamelberg D, Mongan J, McCammon JA: **Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules**. *Journal of Chemical Physics* 2004, **120**:11919-11929.

93. Wereszczynski J, McCammon JA: **Accelerated Molecular Dynamics in Computational Drug Design**. In *Computational Drug Discovery and Design*. Edited by Baron R: Springer New York; 2012:515-524. Methods in Molecular Biology, vol 819.]

94. Gasper PM, Fuglestad B, Komives EA, Markwick PRL, McCammon JA: **Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities**. *Proceedings of the National Academy of Sciences* 2012, **109**:21216-21222.

95. Barducci A, Bonomi M, Parrinello M: **Metadynamics**. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2011, **1**:826-843.

96. Barducci A, Bussi G, Parrinello M: **Well-tempered metadynamics: a smoothly converging and tunable free-energy method**. *Physical Review Letters* 2008, **100**:020603.

97. Smiatek J, Heuer A: **Calculation of free energy landscapes: A histogram reweighted metadynamics approach**. *Journal of Computational Chemistry* 2011, 10.1002/jcc.21790
10.1002/jcc.21790.

98. Tiwary P, Parrinello M: **From Metadynamics to Dynamics**. *Physical Review Letters* 2013, **111**:230602.

99. Tiwary P, Limongelli V, Salvalaglio M, Parrinello M: **Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps**. *Proceedings of the National Academy of Sciences* 2015, **112**:E386-E391.

100. Bonomi M, Parrinello M: **Enhanced Sampling in the Well-Tempered Ensemble**. *Physical Review Letters* 2010, **104**:190601.

101. Palazzesi F, Barducci A, Tollinger M, Parrinello M: **The allosteric communication pathways in KIX domain of CBP**. *Proceedings of the National Academy of Sciences* 2013, **110**:14237-14242.

102. Sugita Y, Okamoto Y: **Replica-exchange molecular dynamics method for protein folding**. *Chemical Physics Letters* 1999, **314**:141-151.

103. Klenin K, Strodel B, Wales DJ, Wenzel W: **Modelling proteins: conformational sampling and reconstruction of folding kinetics**. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2011, **1814**:977-1000.

104. Ishii A, Kimizuka H, Ogata S: **Multi-replica molecular dynamics modeling**. *Computational Materials Science* 2012, **54**:240-248.

105. Pang X, Zhou H-X: **Disorder-to-Order Transition of an Active-Site Loop Mediates the Allosteric Activation of Sortase A**. *Biophysical Journal* **109**:1706-1715.

106. Lane TJ, Shukla D, Beauchamp KA, Pande VS: **To milliseconds and beyond: challenges in the simulation of protein folding**. *Current Opinion in Structural Biology* 2013, **23**:58-65.

107. Harvey MJ, De Fabritiis G: **High-throughput molecular dynamics: the powerful new tool for drug discovery**. *Drug Discovery Today* 2012, **17**:1059-1062.

108. Zwier MC, Chong LT: **Reaching biological timescales with all-atom molecular dynamics simulations**. *Current Opinion in Pharmacology* 2010, **10**:745-752.

109. Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, Shirts MR, Snow CD, Sorin EJ, et al.: **Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing**. *Biopolymers* 2003, **68**:91-109.

110. Bowman GR, Ensign DL, Pande VS: **Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models**. *Journal of Chemical Theory and Computation* 2010, **6**:787-794.

111. Silva D-A, Weiss DR, Pardo Avila F, Da L-T, Levitt M, Wang D, Huang X: **Millisecond dynamics of RNA polymerase II translocation at atomic resolution**. *Proceedings of the National Academy of Sciences* 2014, **111**:7665-7670.

112. Malmstrom RD, Kornev AP, Taylor SS, Amaro RE: **Allostery through the computational microscope: cAMP activation of a canonical signalling domain**. *Nat Commun* 2015, **6**.

113. Shannon CE: **A Mathematical Theory of Communication**. *Bell System Technical Journal* 1948, **27**:379-423.

114. Lange OF, Grubmuller H: **Generalized correlation for biomolecular dynamics**. *Proteins* 2006, **62**:1053-1061.

115. McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP: **Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles**. *Journal of Chemical Theory and Computation* 2009, **5**:2486-2502.

116. Sethi A, Eargle J, Black AA, Luthey-Schulten Z: **Dynamical networks in tRNA:protein complexes**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:6620-6625.

117. Vanwart AT, Eargle J, Luthey-Schulten Z, Amaro RE: **Exploring residue component contributions to dynamical network models of allostery**. *Journal of Chemical Theory and Computation* 2012, **8**:2949-2961.

118. Van Wart AT, Durrant J, Votapka L, Amaro RE: **Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis**. *Journal of Chemical Theory and Computation* 2014, **10**:511-517.

119. Fetics Susan K, Guterres H, Kearney Bradley M, Buhrman G, Ma B, Nussinov R, Mattos C: **Allosteric Effects of the Oncogenic RasQ61L Mutant on Raf-RBD**. *Structure* 2015, **23**:505-516.

120. Pandini A, Fornili A, Fraternali F, Kleinjung J: **Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics**. *The FASEB Journal* 2012, **26**:868-881.

121. Allain A, Chauvot de Beauchene I, Langenfeld F, Guarracino Y, Laine E, Tchertanov L: **Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs**. *Faraday Discussions* 2014, **169**:303-321.

122. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG: **Localizing frustration in native proteins and protein assemblies**. *Proceedings of the National Academy of Sciences* 2007, **104**:19819-19824.

123. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG: **On the role of frustration in the energy landscapes of allosteric proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:3499-3503.

124. Stacklies W, Seifert C, Graeter F: **Implementation of force distribution analysis for molecular dynamics simulations**. *BMC Bioinformatics* 2011, **12**:1-5.

125. Seifert C, Grater F: **Force distribution reveals signal transduction in E. coli Hsp90**. *Biophysical Journal* 2012, **103**:2195-2202.

126. Palmai Z, Seifert C, Gräter F, Balog E: **An Allosteric Signaling Pathway of Human 3-Phosphoglycerate Kinase from Force Distribution Analysis**. *PLoS Computational Biology* 2014, **10**:e1003444.

127. Zhou J, Bronowska A, Le Coq J, Lietha D, Grater F: **Allosteric regulation of focal adhesion kinase by PIP_2 and ATP**. *Biophysical Journal* 2015, **108**:698-705.

128. Stacklies W, Xia F, Gräter F: **Dynamic Allostery in the Methionine Repressor Revealed by Force Distribution Analysis**. *PLoS Computational Biology* 2009, **5**:e1000574.

129. Johnson QR, Lindsay RJ, Nellas RB, Fernandez EJ, Shen T: **Mapping allostery through computational glycine scanning and correlation analysis of residue-residue contacts**. *Biochemistry* 2015, **54**:1534-1541.

130. Soltan Ghoraie L, Burkowski F, Zhu M: **Sparse networks of directly coupled, polymorphic and functional side chains in allosteric proteins**. *Proteins: Structure, Function, and Bioinformatics* 2014, 10.1002/prot.24752:n/a-n/a.

131. LeVine MV, Weinstein H: **NbIT--a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter LeuT**. *PLoS Computational Biology* 2014, **10**:e1003603.

132. Mino-Galaz GA: **Allosteric communication pathways and thermal rectification in PDZ-2 protein: a computational study**. *Journal of Physical Chemistry B* 2015, **119**:6179-6189.

133. Kalescky R, Liu J, Tao P: **Identifying Key Residues for Protein Allostery through Rigid Residue Scan**. *The Journal of Physical Chemistry A* 2015, **119**:1689-1700.

134. Okazaki K-i, Takada S: **Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms**. *Proceedings of the National Academy of Sciences* 2008, **105**:11182-11187.

135. Chakrabarti Kalyan S, Agafonov Roman V, Pontiggia F, Otten R, Higgins Matthew K, Schertler Gebhard FX, Oprian Daniel D, Kern D: **Conformational Selection in a Protein-Protein Interaction Revealed by Dynamic Pathway Analysis**. *Cell Reports* **14**:32-42.

136. Weinkam P, Pons J, Sali A: **Structure-based model of allostery predicts coupling between distant sites**. *Proceedings of the National Academy of Sciences* 2012, **109**:4875-4880.

137. Rodgers TL, Townsend PD, Burnell D, Jones ML, Richards SA, McLeish TCB, Pohl E, Wilson MR, Cann MJ: **Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors**. *PLoS Biology* 2013, **11**:e1001651.

138. McLeish TCB, Rodgers TL, Wilson MR: **Allostery without conformation change: modelling protein dynamics at multiple scales**. *Physical Biology* 2013, **10**:056004.

139. Townsend PD, Rodgers TL, Pohl E, Wilson MR, McLeish TCB, Cann MJ: **Global low-frequency motions in protein allostery: CAP as a model system**. *Biophysical Reviews* 2015, **7**:175-182.

140. Cramer CJ: *Essentials of Computational Chemistry: Theories and Models* edn 2nd: John Wiley & Sons; 2004.

141. Hinchliffe A: *Molecular Modelling for Beginners*: Wiley; 2008.

142. Lonsdale R, Ranaghan KE, Mulholland AJ: **Computational enzymology**. *Chemical Communications* 2010, **46**:2354-2372.

143. Ranaghan KE, Mulholland AJ: **Investigations of enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods**. *International Reviews in Physical Chemistry* 2010, **29**:65-133.

144. Lonsdale R, Harvey JN, Mulholland AJ: **A practical guide to modelling enzyme-catalysed reactions**. *Chemical Society Reviews* 2012, **41**:3025-3038.

145. van der Kamp MW, Mulholland AJ: **Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology**. *Biochemistry* 2013, **52**:2708-2728.

146. Senn H, Thiel W: **QM/MM Methods for Biological Systems**. Edited by Reiher M: Springer Berlin Heidelberg; 2007:173-290. vol 268.]

147. Gao J, Truhlar DG: **Quantum mechanical methods for enzyme kinetics**. *Annual Review of Physical Chemistry* 2002, **53**:467-505.

148. Glowacki DR, Harvey JN, Mulholland AJ: **Taking Ockham's razor to enzyme dynamics and catalysis**. *Nature Chemistry* 2012, **4**:169-176.

149. Swint-Kruse L, Fisher HF: **Enzymatic reaction sequences as coupled multiple traces on a multidimensional landscape**. *Trends in Biochemical Sciences* 2008, **33**:104-112.

150. Benkovic SJ, Hammes GG, Hammes-Schiffer S: **Free-Energy Landscape of Enzyme Catalysis**. *Biochemistry* 2008, **47**:3317-3321.

151. Strickland N, Mulholland AJ, Harvey JN: **The Fe-CO Bond Energy in Myoglobin: A QM/MM Study of the Effect of Tertiary Structure**. *Biophysical Journal* 2006, **90**:L27-L29.

152. Ojeda-May P, Li Y, Ovchinnikov V, Nam K: **Role of Protein Dynamics in Allosteric Control of the Catalytic Phosphoryl Transfer of Insulin Receptor Kinase**. *Journal of the American Chemical Society* 2015, **137**:12454-12457.

153. Harms MJ, Thornton JW: **Evolutionary biochemistry: revealing the historical and physical causes of protein properties**. *Nature Reviews: Genetics* 2013, **14**:559-571.

154. Peracchi A, Mozzarelli A: **Exploring and exploiting allostery: Models, evolution, and drug targeting**. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 2011, **1814**:922-933.

155. Coyle Scott M, Flores J, Lim Wendell A: **Exploitation of Latent Allostery Enables the Evolution of New Modes of MAP Kinase Regulation**. *Cell* 2013, **154**:875-887.

156. Yang J-S, Seo SW, Jang S, Jung GY, Kim S: **Rational Engineering of Enzyme Allosteric Regulation through Sequence Evolution Analysis**. *PLoS Computational Biology* 2012, **8**:e1002612.

157. Lockless SW, Ranganathan R: **Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families**. *Science* 1999, **286**:295-299.

158. Suel GM, Lockless SW, Wall MA, Ranganathan R: **Evolutionarily conserved networks of residues mediate allosteric communication in proteins**. *Nature Structural Biology* 2003, **10**:59-69.

159. de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution**. *Nature Reviews: Genetics* 2013, **14**:249-261.

160. Halabi N, Rivoire O, Leibler S, Ranganathan R: **Protein sectors: evolutionary units of three-dimensional structure**. *Cell* 2009, **138**:774-786.

161. Reynolds Kimberly A, McLaughlin Richard N, Ranganathan R: **Hot Spots for Allosteric Regulation on Protein Surfaces**. *Cell* 2011, **147**:1564-1575.

162. Estabrook RA, Luo J, Purdy MM, Sharma V, Weakliem P, Bruice TC, Reich NO: **Statistical coevolution analysis and molecular dynamics: Identification of amino acid pairs essential for catalysis**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:994-999.

163. Novinec M, Korenč M, Caflisch A, Ranganathan R, Lenarčič B, Baici A: **A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods**. *Nat Commun* 2014, **5**.

164. Bridgham JT, Keay J, Ortlund EA, Thornton JW: **Vestigialization of an Allosteric Switch: Genetic and Structural Mechanisms for the Evolution of Constitutive Activity in a Steroid Hormone Receptor**. *PLoS Genetics* 2014, **10**:e1004058.

165. Jiao W, Hutton RD, Cross PJ, Jameson GB, Parker EJ: **Dynamic Cross-Talk among Remote Binding Sites: The Molecular Basis for Unusual Synergistic Allostery**. *Journal of Molecular Biology* 2012, **415**:716-726.

166. Sprenger GA: **Aromatic Amino Acids**. In *Amino Acid Biosynthesis ~ Pathways, Regulation and Metabolic Engineering*. Edited by Wendisch VF: Springer Berlin Heidelberg; 2007:93-127. 10.1007/7171_2006_067

167. Bentley R: **The Shikimate Pathway — A Metabolic Tree with Many Branches**. *Critical Reviews in Biochemistry and Molecular Biology* 1990, **25**:307-384.

168. Maeda H, Dudareva N: **The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants**. *Annual Review of Plant Biology* 2012, **63**:73-105.

169. Roberts F, Roberts CW, Johnson JJ, Kyle DE, Krell T, Coggins JR, Coombs GH, Milhous WK, Tzipori S, Ferguson DJP, et al.: **Evidence for the shikimate pathway in apicomplexan parasites**. *Nature* 1998, **393**:801-805.

170. Kapnick SM, Zhang Y: **New tuberculosis drug development: targeting the shikimate pathway**. *Expert Opinion on Drug Discovery* 2008, **3**:565-577.

171. Ducati RG, Basso LA, Santos DS: **Mycobacterial shikimate pathway enzymes as targets for drug design**. *Current Drug Targets* 2007, **8**:423-435.

172. Derrer B, Macheroux P, Kappes B: **The shikimate pathway in apicomplexan parasites: Implications for drug development**. *Frontiers in Bioscience-Landmark* 2013, **18**:944-969.

173. Kishore GM, Shah DM: **Amino Acid Biosynthesis Inhibitors as Herbicides**. *Annual Review of Biochemistry* 1988, **57**:627-663.

174. John F: **N-phosphonomethyl-glycine phytotoxicant compositions**. Edited by: Google Patents; 1974.

175. Duke SO, Powles SB: **Glyphosate: a once-in-a-century herbicide**. *Pest Management Science* 2008, **64**:319-325.

176. Light SH, Anderson WF: **The diversity of allosteric controls at the gateway to aromatic amino acid biosynthesis**. *Protein Science* 2013, **22**:395-404.

177. Schofield LR, Anderson BF, Patchett ML, Norris GE, Jameson GB, Parker EJ: **Substrate Ambiguity and Crystal Structure of Pyrococcus furiosus 3-Deoxy-d-arabino-heptulosonate-7-phosphate Synthase: An Ancestral 3-Deoxyald-2-ulosonate-phosphate Synthase?†,‡**. *Biochemistry* 2005, **44**:11950-11962.

178. Zhou L, Wu J, Janakiraman V, Shumilin IA, Bauerle R, Kretsinger RH, Woodard RW: **Structure and characterization of the 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase from Aeropyrum pernix**. *Bioorganic Chemistry* 2012, **40**:79-86.

179. Shumilin IA, Bauerle R, Wu J, Woodard RW, Kretsinger RH: **Crystal Structure of the Reaction Complex of 3-Deoxy-d-arabino-heptulosonate-7-phosphate Synthase from Thermotoga maritima Refines the Catalytic Mechanism and Indicates a New Mechanism of Allosteric Regulation**. *Journal of Molecular Biology* 2004, **341**:455-466.

180. Cross PJ, Dobson RCJ, Patchett ML, Parker EJ: **Tyrosine Latching of a Regulatory Gate Affords Allosteric Control of Aromatic Amino Acid Biosynthesis**. *Journal of Biological Chemistry* 2011, **286**:10216-10224.

181. Light SH, Halavaty AS, Minasov G, Shuvalova L, Anderson WF: **Structural analysis of a 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase with an N-terminal chorismate mutase-like regulatory domain**. *Protein Science* 2012, **21**:887-895.

182. Shumilin IA, Kretsinger RH, Bauerle RH: **Crystal structure of phenylalanine-regulated 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from Escherichia coli**. *Structure* 1999, **7**:865-875.

183. König V, Pfeil A, Braus GH, Schneider TR: **Substrate and Metal Complexes of 3-Deoxy-d-arabino-heptulosonate-7-phosphate Synthase from Saccharomyces cerevisiae Provide New Insights into the Catalytic Mechanism**. *Journal of Molecular Biology* 2004, **337**:675-690.

184. Chaudhury S, Abdulhameed MDM, Singh N, Tawa GJ, D?haeseleer PM, Zemla AT, Navid A, Zhou CE, Franklin MC, Cheung J, et al.: **Rapid Countermeasure Discovery against <italic>Francisella tularensis</italic> Based on a Metabolic Network Reconstruction**. *PloS One* 2013, **8**:e63369.

185. Cross PJ, Pietersma AL, Allison TM, Wilson-Coutts SM, Cochrane FC, Parker EJ: **Neisseria meningitidis expresses a single 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase that is inhibited primarily by phenylalanine**. *Protein Science* 2013, **22**:1087-1099.

186. Webby CJ, Lott JS, Baker HM, Baker EN, Parker EJ: **Crystallization and preliminary X-ray crystallographic analysis of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from Mycobacterium tuberculosis**. *Acta Crystallographica Section F-Structural Biology and Crystallization Communications* 2005, **61**:403-406.

187. Blackmore NJ, Reichau S, Jiao W, Hutton RD, Baker EN, Jameson GB, Parker EJ: **Three Sites and You Are Out: Ternary Synergistic Allostery Controls Aromatic Amino Acid Biosynthesis in Mycobacterium tuberculosis**. *Journal of Molecular Biology* 2013, **425**:1582-1592.

188. Blackmore NJ, Nazmi AR, Hutton RD, Webby MN, Baker EN, Jameson GB, Parker EJ: **Complex Formation between Two Biosynthetic Enzymes Modifies the Allosteric Regulatory Properties of Both: AN EXAMPLE OF MOLECULAR SYMBIOSIS**. *Journal of Biological Chemistry* 2015, **290**:18187-18198.

189. Munack S, Roderer K, Okvist M, Kamarauskaite J, Sasso S, van Eerde A, Kast P, Krengel U: **Remote Control by Inter-Enzyme Allostery: A Novel Paradigm for Regulation of the Shikimate Pathway**. *Journal of Molecular Biology* 2016, 10.1016/j.jmb.2016.01.001.

190. Ahmad S, Rightmire B, Jensen RA: **Evolution of the regulatory isozymes of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase present in the Escherichia coli genealogy**. *Journal of Bacteriology* 1986, **165**:146-154.

191. Hoffmann PJ, Doy CH, Catcheside DEA: **The separation of three allosterically inhibitable 3-deoxy-d-arabino-heptulosonate 7-phosphate synthases from extracts of Neurospora crassa and the purification of the tyrosine inhibitable isoenzyme**. *Biochimica et Biophysica Acta (BBA) - Enzymology* 1972, **268**:550-561.

192. Takahashi M, Chan WWC: **Separation and Properties of Isozymes of 3-Deoxy-D-arabino-heptulosonate-7-phosphate Synthetase from Saccharomyces cerevisiae**. *Canadian Journal of Biochemistry* 1971, **49**:1015-1025.

193. Sousa S, McLaughlin MM, Pereira SA, VanHorn S, Knowlton R, Brown JR, Nicholas RO, Livi GP: **The ARO4 gene of Candida albicans encodes a tyrosine-sensitive DAHP synthase: evolution, functional conservation and phenotype of Aro3p-, Aro4p-deficient mutants**. *Microbiology-Sgm* 2002, **148**:1291-1303.

194. Whitaker RJ, Fiske MJ, Jensen RA: **Pseudomonas aeruginosa possesses two novel regulatory isozymes of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase**. *Journal of Biological Chemistry* 1982, **257**:12789-12794.

195. Walker GE, Dunbar B, Hunter lS, Nimmo HG, Coggins JR: **Evidence for a novel class of microbial 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase in Streptomyces coelicolor A3(2), Streptomyces rimosus and Neurospora crassa**. *Microbiology* 1996, **142**:1973-1982.

196. Silakowski B, Kunze B, Muller R: **Stigmatella aurantiaca Sg a15 carries genes encoding type I and type II 3-deoxy-D-arabino-heptulosonate-7-phosphate synthases: involvement of a type II synthase in aurachin biosynthesis**. *Archives of Microbiology* 2000, **173**:403-411.

197. Henriquez FL, Campbell SJ, Sundararaj BK, Cano A, Muench SP, Roberts CW: **The Acanthamoeba Shikimate Pathway has a Unique Molecular Arrangement and is Essential for Aromatic Amino Acid Biosynthesis**. *Protist* 2015, **166**:93-105.

198. Gosset G, Bonner CA, Jensen RA: **Microbial Origin of Plant-Type 2-Keto-3-Deoxy-d-arabino-Heptulosonate 7-Phosphate Synthases, Exemplified by the Chorismate- and Tryptophan-Regulated Enzyme from Xanthomonas campestris**. *Journal of Bacteriology* 2001, **183**:4061-4070.

199. Robert JG, Thomas AdA, Jensen RA: **The Two-Isozyme System of 3-Deoxy-d-Arabino-Heptulosonate 7-Phosphate Synthase in Nicotiana silvestris and Other Higher Plants**. *Plant Physiology* 1986, **82**:203-210.

200. Keith B, Dong XN, Ausubel FM, Fink GR: **Differential induction of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase genes in Arabidopsis thaliana by wounding and pathogenic attack**. *Proceedings of the National Academy of Sciences* 1991, **88**:8821-8825.

201. Rubin JL, Jensen RA: **Differentially Regulated Isozymes of 3-Deoxy-d-arabino-Heptulosonate-7-Phosphate Synthase from Seedlings of Vigna radiata [L.] Wilczek**. *Plant Physiology* 1985, **79**:711-718.

202. Weaver LM, Herrmann KM: **Dynamics of the shikimate pathway in plants**. *Trends in Plant Science* 1997, **2**:346-351.

203. Herrmann KM, Weaver LM: **THE SHIKIMATE PATHWAY**. *Annual Review of Plant Physiology and Plant Molecular Biology* 1999, **50**:473-503.

204. Herrmann KM: **The Shikimate Pathway as an Entry to Aromatic Secondary Metabolism**. *Plant Physiology* 1995, **107**:7-12.

205. Subramaniam PS, Xie G, Xia T, Jensen RA: **Substrate Ambiguity of 3-Deoxy-d-manno-Octulosonate 8-Phosphate Synthase from Neisseria gonorrhoeae in the Context of Its Membership in a Protein Family Containing a Subset of 3-Deoxy-d-arabino-Heptulosonate 7-Phosphate Synthases**. *Journal of Bacteriology* 1998, **180**:119-127.

206. Birck MR, Woodard RW: **Aquifex aeolicus 3-deoxy-D-manno-2-octulosonic acid 8-phosphate synthase: a new class of KDO 8-P synthase?** *Journal of Molecular Evolution* 2001, **52**:205-214.

207. Jensen R, Xie G, Calhoun D, Bonner C: **The Correct Phylogenetic Relationship of KdsA (3-Deoxy-D-manno-octulosonate 8-Phosphate Synthase) with One of Two Independently Evolved Classes of AroA (3-Deoxy-D-arabino-heptulosonate 7-Phosphate Synthase)**. *Journal of Molecular Evolution* 2002, **54**:416-423.

208. Webby CJ, Baker HM, Lott JS, Baker EN, Parker EJ: **The structure of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase from Mycobacterium tuberculosis reveals a common catalytic scaffold and ancestry for type I and type II enzymes**. *Journal of Molecular Biology* 2005, **354**:927-939.

209. Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW: **Evolutionary Origins of the Eukaryotic Shikimate Pathway: Gene Fusions, Horizontal Gene Transfer, and Endosymbiotic Replacements**. *Eukaryotic Cell* 2006, **5**:1517-1531.

210. Wu J, Woodard RW: **New Insights into the Evolutionary Links Relating to the 3-Deoxy-D-arabino-heptulosonate 7-Phosphate Synthase Subfamilies**. *Journal of Biological Chemistry* 2006, **281**:4042-4048.

211. Zhi X-Y, Yao J-C, Li H-W, Huang Y, Li W-J: **Genome-wide identification, domain architectures and phylogenetic analysis provide new insights into the early evolution of shikimate pathway in prokaryotes**. *Molecular Phylogenetics and Evolution* 2014, **75**:154-164.

212. Angata T, Varki A: **Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective**. *Chemical Reviews* 2002, **102**:439-469.

213. Bravo IG, García-Vallvé S, Romeu A, Reglero Á: **Prokaryotic origin of cytidylyltransferases and α-ketoacid synthases**. *Trends in Microbiology* 2004, **12**:120-128.

214. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al.: **Pfam: the protein families database**. *Nucleic Acids Research* 2014, **42**:D222-D230.

215. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al.: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Molecular Systems Biology* 2011, **7**.

216. Gouy M, Guindon S, Gascuel O: **SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building**. *Molecular Biology and Evolution* 2010, **27**:221-224.

217. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Research* 2011, **39**:W29-W37.

218. Consortium TU: **Activities at the Universal Protein Resource (UniProt)**. *Nucleic Acids Research* 2014, **42**:D191-D198.

219. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al.: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Research* 2011, **39**:D38-D51.

220. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ: **Jalview Version 2—a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**:1189-1191.

221. Frickey T, Lupas A: **CLANS: a Java application for visualizing protein families based on pairwise similarity**. *Bioinformatics* 2004, **20**:3702-3704.

222. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Research* 2004, **32**:1792-1797.

223. Katoh K, Misawa K, Kuma Ki, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Research* 2002, **30**:3059-3066.

224. Notredame C, Higgins DG, Heringa J: **T-coffee: a novel method for fast and accurate multiple sequence alignment1**. *Journal of Molecular Biology* 2000, **302**:205-217.

225. Löytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:10557-10562.

226. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al.: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Research* 2009, **37**:D5-D15.

227. Dress AW, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF: **Noisy: Identification of problematic columns in multiple sequence alignments**. *Algorithms for Molecular Biology* 2008, **3**:1-10.

228. Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C: **Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference**. *Systematic Biology* 2015, **64**:778-791.

229. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009, **25**:1972-1973.

230. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics* 2011, **27**:1164-1165.

231. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**:1312-1313.

232. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: An interactive viewer for large phylogenetic trees**. *BMC Bioinformatics* 2007, **8**:1-6.

233. Rambaut A: **FigTree v1. 3.1: Tree figure drawing tool**. Edited by; 2009.

234. Suchard MA, Redelings BD: **BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny**. *Bioinformatics* 2006, **22**:2047-2048.

235. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: A Sequence Logo Generator**. *Genome Research* 2004, **14**:1188-1190.

236. Nazmi AR, Schofield LR, Dobson RCJ, Jameson GB, Parker EJ: **Destabilization of the Homotetrameric Assembly of 3-Deoxy-d-Arabino-Heptulosonate-7-Phosphate Synthase from the Hyperthermophile Pyrococcus furiosus Enhances Enzymatic Activity**. *Journal of Molecular Biology* 2014, **426**:656-673.

237. Hur S, Bruice TC: **The mechanism of catalysis of the chorismate to prephenate reaction by the Escherichia coli mutase enzyme**. *Proceedings of the National Academy of Sciences* 2002, **99**:1176-1181.

238. Lee AY, Karplus PA, Ganem B, Clardy J: **Atomic structure of the buried catalytic pocket of Escherichia coli chorismate mutase**. *Journal of the American Chemical Society* 1995, **117**:3627-3628.

239. Liu DR, Cload ST, Pastor RM, Schultz PG: **Analysis of Active Site Residues in Escherichia coli Chorismate Mutase by Site-Directed Mutagenesis**. *Journal of the American Chemical Society* 1996, **118**:1789-1790.

240. Lassila JK, Keeffe JR, Kast P, Mayo SL: **Exhaustive Mutagenesis of Six Secondary Active-Site Residues in Escherichia coli Chorismate Mutase Shows the Importance of Hydrophobic Side Chains and a Helix N-Capping Position for Stability and Catalysis†**. *Biochemistry* 2007, **46**:6883-6891.

241. Cross PJ, Allison TM, Dobson RCJ, Jameson GB, Parker EJ: **Engineering allosteric control to an unregulated enzyme by transfer of a regulatory domain**. *Proceedings of the National Academy of Sciences* 2013, **110**:2111-2116.

242. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, et al.: **CATH: comprehensive structural and functional annotations for genome sequences**. *Nucleic Acids Research* 2015, **43**:D376-D381.

243. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I: **New and continuing developments at PROSITE**. *Nucleic Acids Research* 2013, **41**:D344-D347.

244. Drozdetskiy A, Cole C, Procter J, Barton GJ: **JPred4: a protein secondary structure prediction server**. *Nucleic Acids Research* 2015, 10.1093/nar/gkv332.

245. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J: **Template-based protein structure modeling using the RaptorX web server**. *Nat. Protocols* 2012, **7**:1511-1522.

246. Lang EJM, Cross PJ, Mittelstädt G, Jameson GB, Parker EJ: **Allosteric ACTion: the varied ACT domains regulating enzymes of amino-acid metabolism**. *Current Opinion in Structural Biology* 2014, **29**:102-111.

247. Lang EJM, Heyes LC, Jameson GB, Parker EJ: **Calculated pKa Variations Expose Dynamic Allosteric Communication Networks**. *Journal of the American Chemical Society* 2016, **138**:2036-2045.

248. Cross PJ, Heyes LC, Zhang S, Nazmi AR, Parker EJ: **The Functional Unit of Neisseria meningitidis 3-Deoxy--Arabino-Heptulosonate 7-Phosphate Synthase Is Dimeric**. *PloS One* 2016, **11**:e0145187.

249. Martinson VG, Magoc T, Koch H, Salzberg SL, Moran NA: **Genomic Features of a Bumble Bee Symbiont Reflect Its Host Environment**. *Applied and Environmental Microbiology* 2014, **80**:3793-3803.

250. Pible O, Hartmann EM, Imbert G, Armengaud J: **The importance of recognizing and reporting sequence database contamination for proteomics**. *EuPA Open Proteomics* 2014, **3**:246-249.

251. Redelings B: **Erasing Errors due to Alignment Ambiguity When Estimating Positive Selection**. *Molecular Biology and Evolution* 2014, **31**:1979-1993.

252. McKenzie SK, Oxley PR, Kronauer DJ: **Comparative genomics and transcriptomics in ants provide new insights into the evolution and function of odorant binding and chemosensory proteins**. *BMC Genomics* 2014, **15**:1-14.

253. Ahn M, Pietersma AL, Schofield LR, Parker EJ: **Mechanistic divergence of two closely related aldol-like enzyme-catalysed reactions**. *Organic & Biomolecular Chemistry* 2005, **3**:4046-4049.

254. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation**. *Nat Biotech* 2012, **30**:1072-1080.

255. Wu J, Howe DL, Woodard RW: **Thermotoga maritima 3-Deoxy-D-arabino-heptulosonate 7-Phosphate (DAHP) Synthase: THE ANCESTRAL EUBACTERIAL DAHP SYNTHASE?** *Journal of Biological Chemistry* 2003, **278**:27525-27531.

256. Cross PJ, Parker EJ: **Unravelling the Evolution of Allosteric Regulation in 3-Deoxy-D-arabino-heptulosonate 7-phosphate Synthase**: Univerity of Canterbury: 2012.

257. Cross PJ, Parker EJ: **Allosteric inhibitor specificity of Thermotoga maritima 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase**. *FEBS Letters* 2013, **587**:3063-3068.

258. Emsley P, Cowtan K: **Coot: model-building tools for molecular graphics**. *Acta Crystallographica Section D* 2004, **60**:2126-2132.

259. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *Journal of Molecular Biology* 1993, **234**:779-815.

260. Song Y, Mao J, Gunner MR: **MCCE2: Improving protein pKa calculations with extensive side chain rotamer sampling**. *Journal of Computational Chemistry* 2009, **30**:2231-2247.

261. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH: **PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions**. *Journal of Chemical Theory and Computation* 2011, **7**:525-537.

262. Mackerell AD, Jr., Feig M, Brooks CL, 3rd: **Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in**

reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* 2004, **25**:1400-1415.

263. Tikhonova IG, Selvam B, Ivetac A, Wereszczynski J, McCammon JA: **Simulations of Biased Agonists in the β2 Adrenergic Receptor with Accelerated Molecular Dynamics**. *Biochemistry* 2013, **52**:5593-5603.

264. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, et al.: **GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit**. *Bioinformatics* 2013, **29**:845-854.

265. Humphrey W, Dalke A, Schulten K: **VMD: Visual molecular dynamics**. *Journal of Molecular Graphics* 1996, **14**:33-38.

266. Oliveira SH, Ferraz FA, Honorato RV, Xavier-Neto J, Sobreira TJ, de Oliveira PS: **KVFinder: steered identification of protein cavities as a PyMOL plugin**. *BMC Bioinformatics* 2014, **15**:1-8.

267. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility**. *Journal of Computational Chemistry* 2009, **30**:2785-2791.

268. Trott O, Olson AJ: **AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *Journal of Computational Chemistry* 2010, **31**:455-461.

269. Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB, Stetter KO: **Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C**. *Archives of Microbiology* **144**:324-333.

270. Šali A, Blundell TL: **Comparative Protein Modelling by Satisfaction of Spatial Restraints**. *Journal of Molecular Biology* 1993, **234**:779-815.

271. Schrodinger, LLC: **The PyMOL Molecular Graphics System, Version 1.3r1**. Edited by; 2010.

272. Zoete V, Cuendet MA, Grosdidier A, Michielin O: **SwissParam: A fast force field generation tool for small organic molecules**. *Journal of Computational Chemistry* 2011, **32**:2359-2368.

273. Allinger NL, Yuh YH, Lii JH: **Molecular mechanics. The MM3 force field for hydrocarbons. 1**. *Journal of the American Chemical Society* 1989, **111**:8551-8566.

274. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics**. *Journal of the American Chemical Society* 1989, **111**:8566-8575.

275. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons**. *Journal of the American Chemical Society* 1989, **111**:8576-8582.

276. Fiser A, Do RKG, Šali A: **Modeling of loops in protein structures**. *Protein Science* 2000, **9**:1753-1773.

277. Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV, Svergun DI: **New developments in the ATSAS program package for small-angle scattering data analysis**. *Journal of Applied Crystallography* 2012, **45**:342-350.

278. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R: **Analysis Tool Web Services from the EMBL-EBI**. *Nucleic Acids Research* 2013, **41**:W597-W600.

279. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends in Genetics* **16**:276-277.

280. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, Taly J-F, Notredame C: **T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension**. *Nucleic Acids Research* 2011, **39**:W13-W17.

281. Robert X, Gouet P: **Deciphering key features in protein structures with the new ENDscript server**. *Nucleic Acids Research* 2014, **42**:W320-W324.

282. Rambo RP, Tainer JA: **Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law**. *Biopolymers* 2011, **95**:559-571.

283. Hammel M: **Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS)**. *European Biophysics Journal* 2012, **41**:789-799.

284. Mertens HD, Svergun DI: **Structural characterization of proteins and complexes using small-angle X-ray solution scattering**. *Journal of Structural Biology* 2010, **172**:128-141.

285. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, et al.: **CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields**. *Journal of Computational Chemistry* 2010, **31**:671-690.

286. Mayne CG, Saam J, Schulten K, Tajkhorshid E, Gumbart JC: **Rapid parameterization of small molecules using the force field toolkit**. *Journal of Computational Chemistry* 2013, **34**:2757-2770.

287. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, et al.: **Gaussian 09**. Edited by. Wallingford, CT, USA: Gaussian, Inc.; 2009.

288. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering**. *Journal of the American Chemical Society* 2007, **129**:5656-5664.

289. Pelikan M, Hura GL, Hammel M: **Structure and flexibility within proteins as identified through small angle X-ray scattering**. *General Physiology and Biophysics* 2009, **28**:174-189.

290. Gu S, Silva D-A, Meng L, Yue A, Huang X: **Quantitatively Characterizing the Ligand Binding Mechanisms of Choline Binding Protein Using Markov State Model Analysis**. *PLoS Computational Biology* 2014, **10**:e1003767.

291. Hammes GG, Chang Y-C, Oas TG: **Conformational selection or induced fit: A flux description of reaction mechanism**. *Proceedings of the National Academy of Sciences* 2009, **106**:13737-13741.

292. Case D, Babin V, Berryman J, Betz R, Cai Q, Cerutti D, Cheatham III T, Darden T, Duke R, Gohlke H: **Amber 14**. 2014.

293. Reichau S, Jiao W, Walker SR, Hutton RD, Baker EN, Parker EJ: **Potent Inhibitors of a Shikimate Pathway Enzyme from Mycobacterium tuberculosis: COMBINING MECHANISM- AND MODELING-BASED DESIGN**. *Journal of Biological Chemistry* 2011, **286**:16197-16207.

294. Heyes LC, Reichau S, Cross PJ, Jameson GB, Parker EJ: **Structural analysis of substrate-mimicking inhibitors in complex with Neisseria meningitidis 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase - The importance of accommodating the active site water**. *Bioorganic Chemistry* 2014, **57**:242-250.

295. Tao P, Gatti DL, Schlegel HB: **The energy landscape of 3-deoxy-D-manno-octulosonate 8-phosphate synthase**. *Biochemistry* 2009, **48**:11706-11714.

296. Salomon-Ferrer R, Case DA, Walker RC: **An overview of the Amber biomolecular simulation package**. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2013, **3**:198-210.

297. Götz AW, Clark MA, Walker RC: **An extensible interface for QM/MM molecular dynamics simulations with amber**. *Journal of Computational Chemistry* 2013, 10.1002/jcc.23444:n/a-n/a.

298. Grubmüller H, Groll V: **SOLVATE v1.0** Edited by. Theoretical Biophysics Group, Institute for Medical Optics, Ludwig-Maximilians University, Munich; 1996.

299. Crowley MF, Williamson MJ, Walker RC: **CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software**. *International Journal of Quantum Chemistry* 2009, **109**:3767-3772.

300. DeLeo AB, Dayan J, Sprinson DB: **Purification and Kinetics of Tyrosine-sensitive 3-Deoxy-d-arabino-heptulosonic Acid 7-Phosphate Synthetase from Salmonella**. *Journal of Biological Chemistry* 1973, **248**:2344-2353.

301. Hedstrom L, Abeles R: **3-Deoxy-D-manno-octulosonate-8-phosphate synthase catalyzes the C-O bond cleavage of phosphoenolpyruvate**. *Biochemical and Biophysical Research Communications* 1988, **157**:816-820.

302. Gunawan J, Simard D, Gilbert M, Lovering AL, Wakarchuk WW, Tanner ME, Strynadka NCJ: **Structural and Mechanistic Analysis of Sialic Acid Synthase NeuB from Neisseria meningitidis in Complex with Mn2+, Phosphoenolpyruvate, and N-Acetylmannosaminitol**. *Journal of Biological Chemistry* 2005, **280**:3555-3563.

303. Lopez X, York DM: **Parameterization of semiempirical methods to treat nucleophilic attacks to biological phosphates: AM1/d parameters for phosphorus**. *Theoretical Chemistry Accounts* 2003, **109**:149-159.

304. Nam K, Cui Q, Gao J, York DM: **Specific Reaction Parametrization of the AM1/d Hamiltonian for Phosphoryl Transfer Reactions: H, O, and P Atoms**. *Journal of Chemical Theory and Computation* 2007, **3**:486-504.

305. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA: **THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method**. *Journal of Computational Chemistry* 1992, **13**:1011-1021.

306. Bertini I: *Biological inorganic chemistry: structure and reactivity*: University Science Books; 2007.

307. Hagen WR: **EPR spectroscopy as a probe of metal centres in biological systems**. *Dalton Transactions* 2006, 10.1039/B608163K:4415-4434.

308. Osorio EA, Moth-Poulsen K, van der Zant HSJ, Paaske J, Hedegård P, Flensberg K, Bendix J, Bjørnholm T: **Electrical Manipulation of Spin States in a Single Electrostatically Gated Transition-Metal Complex**. *Nano Letters* 2010, **10**:105-110.

309. Liao R-Z, Thiel W: **Determinants of Regioselectivity and Chemoselectivity in Fosfomycin Resistance Protein FosA from QM/MM Calculations**. *The Journal of Physical Chemistry B* 2013, **117**:1326-1336.

310. Metz S, Kästner J, Sokol AA, Keal TW, Sherwood P: **ChemShell—a modular software package for QM/MM simulations**. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2014, **4**:101-110.

311. Todorov IT, Smith W, Trachenko K, Dove MT: **DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism**. *Journal of Materials Chemistry* 2006, **16**:1911-1918.

312. Neese F: **The ORCA program system**. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2012, **2**:73-78.

313. Rommel JB, Kästner J: **The Fragmentation–Recombination Mechanism of the Enzyme Glutamate Mutase Studied by QM/MM Simulations**. *Journal of the American Chemical Society* 2011, **133**:10195-10203.

314. Dudev T, Lim C: **Factors Governing the Protonation State of Cysteines in Proteins: An Ab Initio/CDM Study**. *Journal of the American Chemical Society* 2002, **124**:6759-6766.

315. Marenich AV, Cramer CJ, Truhlar DG: **Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions**. *The Journal of Physical Chemistry B* 2009, **113**:6378-6396.

316. Vriend G: **WHAT IF: A molecular modeling and drug design program**. *Journal of Molecular Graphics* 1990, **8**:52-56.

317. Li P, Roberts BP, Chakravorty DK, Merz KM: **Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent**. *Journal of Chemical Theory and Computation* 2013, **9**:2733-2748.

318. Yesselman JD, Price DJ, Knight JL, Brooks CL: **MATCH: An atom-typing toolset for molecular mechanics force fields**. *Journal of Computational Chemistry* 2012, **33**:189-202.

319. Foloppe N, Sagemark J, Nordstrand K, Berndt KD, Nilsson L: **Structure, dynamics and electrostatics of the active site of glutaredoxin 3 from Escherichia coli: comparison with functionally related proteins1**. *Journal of Molecular Biology* 2001, **310**:449-470.

320. Cao L, Dong G, Lai W: **Reaction Mechanism of Cobalt-Substituted Homoprotocatechuate 2,3-Dioxygenase: A QM/MM Study**. *The Journal of Physical Chemistry B* 2015, **119**:4608-4616.

321. Luo S, Averkiev B, Yang KR, Xu X, Truhlar DG: **Density Functional Theory of Open-Shell Systems. The 3d-Series Transition-Metal Atoms and Their Cations**. *Journal of Chemical Theory and Computation* 2014, **10**:102-121.

322. Elliott TS, Slowey A, Ye Y, Conway SJ: **The use of phosphate bioisosteres in medicinal chemistry and chemical biology**. *MedChemComm* 2012, **3**:735-751.